Astronomy & Astrophysics manuscript no. NetZ
November 26, 2020

Photometric Redshift Estimation with a Convolutional Neural Network: NetZ

S. Schuldt[1,2], S. H. Suyu[1,2,3], R. Cañameras[1], S. Taubenberger[1], T. Meinhard[4], L. Leal-Taixé[4], and B.C. Hsieh[3]

[1] Max-Planck-Institut für Astrophysik, Karl-Schwarzschild Str. 1, 85741 Garching, Germany
e-mail: schuldt@mpa-garching.mpg.de
[2] Physik Department, Technische Universität München, James-Franck Str. 1, 85741 Garching, Germany
[3] Institute of Astronomy and Astrophysics, Academia Sinica, 11F of ASMAB, No.1, Section 4, Roosevelt Road, Taipei 10617, Taiwan
[4] Informatik Department, Technische Universität München, Bolzmannstr. 3, 85741 Garching, Germany

Received –; accepted –

ABSTRACT

The redshifts of galaxies are a key attribute that is needed for nearly all extragalactic studies. Since spectroscopic redshifts require additional telescope and human resources, millions of galaxies are known without spectroscopic redshifts. Therefore, it is crucial to have methods for estimating the redshift of a galaxy based on its photometric properties, the so-called photo-z. We developed NetZ, a new method using a Convolutional Neural Network (CNN) to predict the photo-z based on galaxy images, in contrast to previous methods which often used only the integrated photometry of galaxies without their images. We use data from the Hyper Suprime-Cam Subaru Strategic Program (HSC SSP) in five different filters as training data. The network over the whole redshift range between 0 and 4 performs well overall and especially in the high-z range better than other methods on the same data. We obtain an accuracy $|z_{\rm pred} - z_{\rm ref}|$ of $\sigma = 0.12$ (68% confidence interval) with a CNN working for all galaxy types averaged over all galaxies in the redshift range of 0 to ∼4. By limiting to smaller redshift ranges or to Luminous Red Galaxies (LRGs), we find a further notable improvement. We publish more than 34 million new photo-z values predicted with NetZ here. This shows that the new method is very simple and fast to apply, and, importantly, covers a wide redshift range limited only by the available training data. It is broadly applicable and beneficial to imaging surveys, particularly upcoming surveys like the Rubin Observatory Legacy Survey of Space and Time which will provide images of billions of galaxies with similar image quality as HSC.

Key words. methods: data analysis, surveys, galaxies: distances and redshifts, galaxies: general, cosmology: observations

1. Introduction

Past imaging surveys have detected billions of galaxies over the sky, a number that will grow significantly with forthcoming wide field surveys such as the Rubin Observatory Legacy Survey in Space and Time (LSST). In most applications for which galaxies are used, redshifts are needed, but spectroscopic redshifts are available only for a small fraction of them. Therefore photometric redshift techniques (hereafter photo-z, see Hildebrandt et al. 2010, and references therein) were developed and improved over the last decades (e.g., Coupon et al. 2009; Hildebrandt et al. 2008, 2012; Dahlen et al. 2013; Bonnett et al. 2016; Tanaka et al. 2018). Typically, photometry in multiple wavelength bands is used to minimize the difference between spectroscopically confirmed redshifts and the predicted photometric redshifts.

There are nowadays two main families of photo-z methods, template fitting and machine learning (ML) methods. They are complementary and both capable of predicting very precise photo-z. Template fitting codes (e.g., Arnouts et al. 1999; Bolzonella et al. 2000; Feldmann et al. 2006; Brammer et al. 2008) are mainly based on galaxy spectral energy distribution (SED) template libraries. This method is physically motivated and already well studied. The templates are used to match the observed colors with the predicted ones (through so-called nearest neighbor algorithms). Such an approach represents the opportunity to provide photo-z estimates in regions of color/magnitude space where no reference redshifts are available. ML provides another approach to get very precise and fast photo-z estimates (e.g., Tagliaferri et al. 2003; Collister & Lahav 2004; Lima et al. 2008; Wolf 2009; Carliles et al. 2010; Singal et al. 2011; Hoyle 2016; Tanaka et al. 2018; Bonnett 2015; D'Isanto & Polsterer 2018; Eriksen et al. 2020; Schmidt et al. 2020). The main requirement is a training sample with known (i.e. spectroscopic or very good photo-z) reference redshifts, which should match the expected redshift distribution. Depending on the network architecture, ML codes in general look for specific features in the training sample and try to extract the important information. So far most algorithms are based on photometric parameters like color/magnitude measurements or also size/compactness measurements and often limited to a narrow redshift range, e.g., up to $z = 1$ (e.g., Bonnett 2015; Hoyle 2016; Sadeh et al. 2016; Almosallam et al. 2016b; Pasquet-Itam & Pasquet 2018; Pasquet et al. 2019; Eriksen et al. 2020; Campagne 2020).

With the success of CNNs in image processing, we investigate a network which estimates photo-z based directly on images of galaxies. This is similar to the work done by Hoyle (2016), where images of galaxies are converted into magnitude images and pixel color maps to feed the architecture, but our network accepts the images directly as observed. Moreover, while Hoyle (2016) use a classification network where the galaxies are sorted into redshift bins between 0 and 1, we use a regression network. This means our network predicts one specific number for the galaxy redshift. The work presented by D'Isanto & Polsterer (2018) and Pasquet et al. (2019) explores both networks with

CNN layers on SDSS galaxies to obtain a probability density function (PDF). D'Isanto & Polsterer (2018) have tested networks for either quasars or galaxies as well as a combination of stars, quasars and galaxies. For the galaxy sample they limit to $0 < z < 1$, although most galaxies are at the lower end such that Pasquet et al. (2019) directly limit the range up to $z = 0.4$. In comparison to those two networks, we have many more galaxies with higher redshifts ($z \sim 1 - 3$) and thus do not limit the redshift range in order to obtain a more powerful network which is directly applicable to the expected redshift range covered by LSST. Based on the available reference redshifts, we test the performance up to a redshift of 4. Since we provide images of different filters, our CNN is able to extract the color and magnitude parameters internally and output a photo-z value at the end. It is trained on images observed in 5 different filters, explicitly on Hyper Suprime-Cam Subaru Strategic Program (HSC SSP, hereafter HSC; Aihara et al. 2018) *grizy* images of galaxies with known spectroscopic or reliable ~30-band photometric redshifts.

The outline of the paper is as follows. We describe in Sec. 2 the training data we are using and we give a short introduction and overview of the used network architecture in Sec. 3. Our main network is presented in Sec. 4 and we compare our NetZ results to other model techniques in Sec. 5. We show in Section 6 our results of the network specialized for LRGs and low redshift range, while we summarize our results in Sec. 7.

2. Training Data

We use images from PDR2 of the HSC-SSP[1] survey (Aihara et al. 2019) for the training of the CNN. The HSC is a wide-field optical camera with a field of view of 1.8 square degrees (1.5 degree in diameter) installed at the 8.2m Subaru Telescope. The data release covers over 300 square degrees of the night sky in five optical filter, *grizy*. The exposure time is 10 minutes for the filters g and r, and 20 minutes for i, z, and y, yielding limiting magnitudes around 26. The pixel size is 0.168″ such that our cutouts with 64×64 pixels result in images of around $10'' \times 10''$. The median seeing in the i-band is 0.6″.

The catalog of all available galaxies from HSC PDR2 in the wide area that pass the criteria

- {*grizy*}_cmodel_flux_flag is *False*
- {*grizy*}_pixelflags_edge is *False*
- {*grizy*}_pixelflags_interpolatedcenter is *False*
- {*grizy*}_pixelflags_saturatedcenter is *False*
- {*grizy*}_pixelflags_crcenter is *False*
- {*grizy*}_pixelflags_bad is *False*
- {*grizy*}_sdsscentroid_flag is *False*

includes around 190 Million galaxies and is represented by a green box in Figure 1. The corresponding HSC images can be used as input data for the network NetZ.

As ground truth we use the spectroscopic redshifts provided by the HSC team, which is a collection from various spectroscopic surveys (zCOSMOS DR3 (Lilly et al. 2009), UDSz (Bradshaw et al. 2013; McLure et al. 2012), 3D-HST (Skelton et al. 2014; Momcheva et al. 2016), VVDS (Le Fèvre et al. 2013), VIPERS PDR1 (Garilli et al. 2014), SDSS DR14 (Alam et al. 2015), GAMA DR2 (Liske et al. 2015), DEEP3 (Davis et al. 2003; Newman et al. 2013), PRIMUS DR1 (Coil et al. 2011; Cool et al. 2013)) where we exclude from SDSS the objects from BOSS/eBOSS and quasars as well as WiggleZ (Drinkwater et al. 2010) and FMOS-COSMOS (Silverman et al. 2015) data since these surveys focused on specific galaxy types/features and would bias our sample. This spec-z sample is cleaned with the following criteria:

[1] HSC webpage: https://hsc-release.mtk.nao.ac.jp/doc/

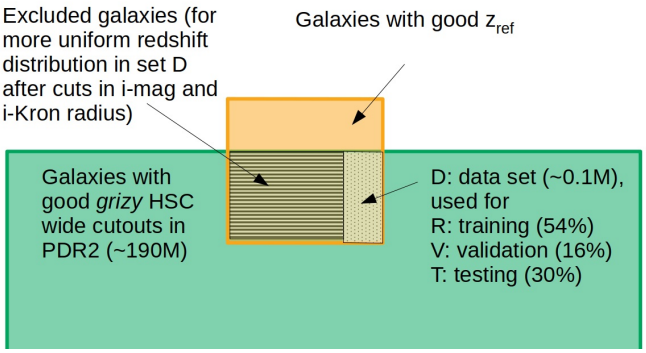

Fig. 1: Sketch of the available data and the intersection of the data D (dotted) used for training (R), validation (V) and testing (T) of the main network $\mathrm{CNN}_{\mathrm{main}}$ presented in Section 4.

- source type is *GALAXY* or *LRG*
- $z > 0$
- $z \neq 9.9999$[2]
- $0 < z_{\mathrm{err}} < 1$
- the galaxy identification number (*ID*) is unique
- specz_flag_homogeneous is *False* (homogenized spec-z flag from HSC team)

This spec-z sample is used in combination with COSMOS2015 (Laigle et al. 2018), a photo-z catalog of the COSMOS field based on around 30 available filters, where we enforce the following criteria:

- flag_capak is *False*
- type = 0 (only galaxies)
- $\chi^2(\mathrm{gal}) < \chi^2(\mathrm{star})$ and $\chi^2(\mathrm{gal})/\mathrm{Nbands} < 5$ (fits are reasonable and better than stellar alternatives)
- ZP_2< 0 (no secondary peak)
- $\log(M_\star) > 7.5$ (stellar mass successfully recovered)
- $0 < z < 9$
- $\max(z_{84} - z_{50}, z_{50} - z_{16}) < 0.05(1 + z)$ (1σ-redshift dispersion $< 5\%$)

The combined cleaned catalog is shown as a yellow box in Figure 1, and the overlap with available good HSC images in all five filter (green box) contains 406,540 galaxies. We further select galaxies with i-band magnitude brighter than 25 mag and a Kron radius larger than 0.8″ in the i band. The limit on the Kron radius is chosen to obtain a set that is representing best the sample we apply NetZ to.

Based on various tests during the development, we found a significant improvement by masking the background and surrounding objects next to the galaxy of interest with Source Extractor (Bertin & Arnouts 1996) before feeding them into the CNN. As boundary we use the 3σ level of the background. Neighboring objects in the field are excluded by requesting the object center to be within 5 pixels of the image center. With this method, we keep only the central galaxy(ies) in the image cutout.

[2] This is the upper limit of the catalog and thus treated as no spec-z available, i.e. excluded

At the end we convolve the extracted image with a gaussian kernel of size 3×3 pixels and a width of 1.5 pixels to smooth out the boundaries very slightly. For those galaxies passing our criteria, we find a better accuracy for the bright galaxies than fainter galaxies.

Since the distribution of the reference redshifts $z_{\rm ref}$ in the training set is very important for the network, we limit each redshift bin of width 0.01 to have no more than 1000 galaxies from those passing the above criteria. This essentially limits the number of low-redshift galaxies which would otherwise be overrepresented in the training set. As a result, the redshift distribution becomes more uniform and allows the CNN to learn/predict redshifts for the full redshift range rather than only the lower-redshift end. The excluded galaxy sample is marked in the underlying yellow box with lines while the sample D is used for our main network NetZ$_{\rm main}$.

Since the sample D is still dominated by galaxies with $z_{\rm ref} < 1$, we test the effect of data augmentation. Explicitly, we include rotated images for $z_{\rm ref} > 1$ and in addition mirrored images for $z_{\rm ref} > 2$ and thus obtain a nearly uniform distribution up to $z_{\rm ref} \sim 1.5$ (see histogram in Figure 2).

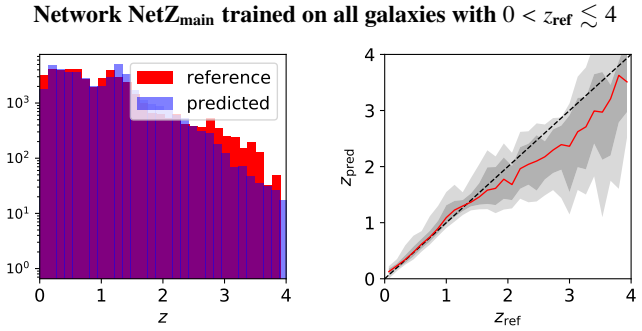

Fig. 2: Performance of the final network on the test set T. On the left hand side, histograms of the reference and predicted redshift distributions are shown in red and blue, respectively. On the right hand side, a 1:1 comparison of reference and predicted redshifts is plotted. The red line shows the median predicted redshift per bin and the gray bands the 1σ and 2σ confidence levels. While the red line follows the black dashed reference line for low redshift very nicely, NetZ$_{\rm main}$ tends to underpredict the high end.

3. Deep learning and the network architecture

Neural Networks (NN) are very powerful tools for many different tasks and especially for work with a huge amount of data. Substantial effort has therefore been spent on deep learning (DL) developments in the recent years. In general, for supervised learning one needs a data set where the input and output, the so-called ground truth, are known. On this data, the network is trained and can afterwards be applied to new data where the output is not known. The main advantages of NN are the variety of architectures and thus the broad range of problem they can be applied to, as well as the speed of those networks in comparison to other methods. Generally, there are two kinds of networks; classification networks distinguish between different classes of objects whereas regression networks predict specific numerical quantities. The latter is the kind of network we are using here, i.e., the network shall predict a specific value of a galaxy's redshift.

Depending on the task, there are different types of networks. Since our input consists of images of galaxies, a typical type is the CNN where the fully-connected (FC) layers are preceded by a number of convolutional (conv) layers. The detailed architecture depends on different parameters like the specific task, the size of the images, and the size of the data set. We tested different architectures and found an overall good network behavior with two convolutional layers followed by three FC layers. We tested different constructions of CNN architectures by varying the number of convolutional or FC layers, strides and kernel sizes but found no improvement. A sketch of the final architecture is shown in Figure 3. The input consists of five different filters for each galaxy and each image has a size of 64×64 pixels, corresponding to an image size of around $10'' \times 10''$. The convolutional layers have stride $s = 1$ and a kernel size of $5 \times 5 \times C$, where $C = 5$ in the first convolutional layer, and $C = 32$ in the second layer. We use 32 kernels and 64 kernels in the first and second convolutional layers, respectively. Each convolutional layer is followed by max pooling of size 2×2 and of stride 2. This results in a data cube of size $13\times 13 \times 64$ which after flattening is passed on to the FC layers to obtain the single output value, the redshift of the galaxy.

Independent of the network architecture, the network can contain hundreds of thousands (or more) of neurons. While at the beginning the values of the weight parameters and bias of each neuron are random, they will be updated at every iteration of the training. To see the network performance after the training, one splits the data into three sets, the training set R, the validation set V and the test set T (see Figure 1). In our case we used 56% of the data set as training set, 14% as validation set, and 30% as test set. We train over 300 epochs, and divide each epoch into a number of iterations by splitting the training, validation, and test set into batches of size N. In each iteration a batch is passed through the entire network to predict the redshifts $z_{\rm pred}$ (forward propagation). The difference between those predicted values and the ground truth is quantified by the loss function L, where we use the mean-square-error (MSE) defined as [3]

$$L = \frac{1}{N}\sum_{k=1}^{N}(z_{{\rm pred},k} - z_{{\rm ref},k})^2 \,. \qquad (1)$$

After the forward propagation and computing the loss for the batch, the information is propagated to the weights and biases (back propagation) which are then modified using a stochastic gradient descent algorithm with a momentum of 0.9. This procedure is repeated for all batches in the training set and a total training loss for this epoch is obtained. Afterwards, the loss is computed within the validation set to determine the improvement of the network, which concludes the epoch.

We perform a so-called cross validation to minimize bias in the validation set, which means to train the NN on the training set and use the validation set to validate the performance after each epoch as described above. These steps are repeated by exchanging the validation set within the training set, such that we have with our splitting five cross-validation runs. At the end, the network is trained on training and validation set together and terminated at the best epoch of all cross validation runs. The best epoch is defined as the epoch with the minimal average vali-

[3] This definition is for only one parameter, which in our case is the redshift. For a general expression, one would also sum over the different parameters.

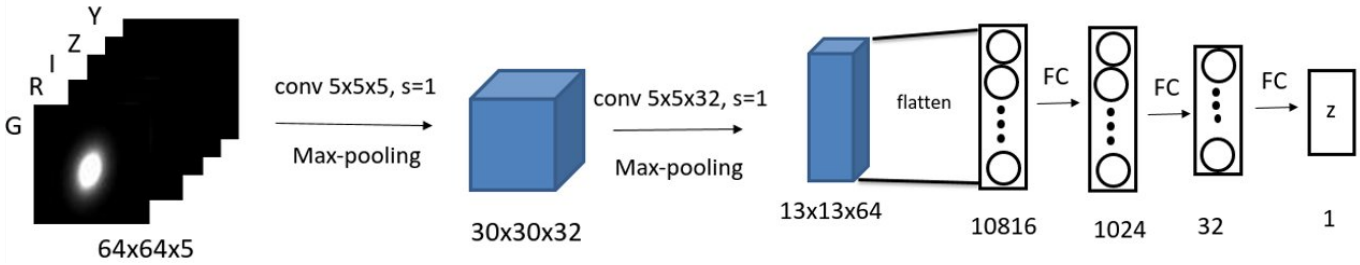

Fig. 3: Overview of the CNN architecture. It contains two convolutional (conv) layers with max pooling and three Fully Connected (FC) layers. The input consists of images of size 64×64 pixels in five different filters *grizy*. The output is the predicted photometric redshift.

dation loss. This network is then applied to the test set, which contains data the network has never seen before.

4. Main Redshift Network NetZ$_{\rm main}$

In this section, we present our main network NetZ$_{\rm main}$ which is trained in the full redshift range ($0 < z \lesssim 4$). We find that this CNN is overall very precise in predicting redshifts. Figure 2 shows a comparison of our final network predictions $z_{\rm pred}$ to the reference redshifts $z_{\rm ref}$ of the test set T. In detail, the left panel of this plot shows a histogram of the reference redshifts (red) and predicted values (blue). On the right panel, a 1:1 comparison of reference and predicted redshifts is plotted. The red line shows the median and the gray bands the 1σ and 2σ confidence levels.

While the network performance is good in the redshift range between 0 and ∼2, the network starts to underestimate the higher redshifts. This is understandable as the network is trained on many more images in the lower reshift range as we can see directly from the histogram. The reason is the limited amount of available training data (reference redshifts) above $z \sim 2$. Moreover, these distant galaxies are typically faint and small in extent, which complicates the learning of their morphological features.

As described in Section 3, we use cross-validation and train always over 300 epochs. We do not see much overfitting from the loss curve, where overfitting means that after a certain number of epochs the network learns to predict the redshifts better for the training set than for the validation set. Based on our testing of different hyper-parameters such as batch size or learning rate, the best moment for terminating the training of NetZ$_{\rm main}$ is at epoch 135 with a loss of 0.1107 according to the loss function L. This network has a learning rate of 0.0005 and a batch size of 128. We also tested drop-out, which means to ignore during each training epoch a new random set of neurons. This can help to reduce overfitting and balance the importance of the neurons in the network. We tested with a dropout rate of 0.5 between the FC layers, but it turned out that drop-out is not necessary for this network.

We test the network performance by varying the masking, especially the kernel for the smoothing. The difference of $\lesssim 0.01$ in the predictions is small compared to the typical photo-z uncertainty (as we see in the scatter of Figure 2). This network stability is important in case the extraction is not done perfectly as planned and done for the training. The masking is done in the exact same way for the newly predicted photo-z values as for training and testing.

It turns out that the network predicts similar but slightly different values for the augmented images, which shows that the network does not identify the rotated/mirrored images as duplicates. The possibility to use such data augmentation and hence boost the performance at high redshifts is a major strength of NetZ.

5. Comparison of NetZ$_{\rm main}$ to other photo-z methods

5.1. Detailed comparison to HSC method DEmP

Since there are already different photo-z methods developed and applied to the HSC data, we show here a comparison. It is very important to use the same data set for a fair comparison. Thus, we only compare to the DEmP (Hsieh & Yee 2014) method where we have a predicted photo-z value for each galaxy within our test set T, and have used identical training and validation sets as for NetZ$_{\rm main}$ without data augmentation as DEmP also relies on the reference distribution. DEmP is one of the best-performing methods from the HSC photo-z team (Tanaka et al. 2018; Nishizawa et al. 2020) and thus a good performance benchmark. DEmP is a hybrid photo-z method by combining polynomial fitting and a N-nearest neighbor method based on photometric values on a catalog level. Therefore, the input data are totally different from those of NetZ, which is based on the pixelated image cutouts of the galaxies. For the comparison, we adopt three quantities from the HSC photo-z papers (Tanaka et al. 2018; Nishizawa et al. 2020), which are defined as follows for each redshift bin

- **Bias** $$\text{Median}\,(\Delta z_i) = \text{Median}\left(\frac{z_{\text{pred},i} - z_{\text{ref},i}}{1 + z_{\text{ref},i}}\right) \quad (2)$$
- **Dispersion** $$\sigma = 1.48 \times \text{MAD}(\Delta z_i) = \\ = 1.48 \times \text{Median}\,(|\Delta z_i - \text{Median}(\Delta z_i)|) \quad (3)$$
- **Outlier rate** $$f_{\text{outlier}} = \frac{N(|\Delta z_i| > 0.15)}{N_{\text{bin}}} \quad (4)$$

where i denotes the i^{th} galaxy in the redshift bin, $z_{\rm pred}$ the predicted photometric redshift, $z_{\rm ref}$ the reference redshift, N the number of galaxies satisfying the specified condition, and $N_{\rm bin}$ the total number of galaxies in the bin. The dispersion is defined using the Median Absolute Deviation (MAD) as expressed above. The multiplication factor comes from statistics and is the relation factor for normally distributed data between MAD and the standard deviation (Rousseeuw & Croux 1993). The comparison is shown in Figure 4, in black triangles the performance

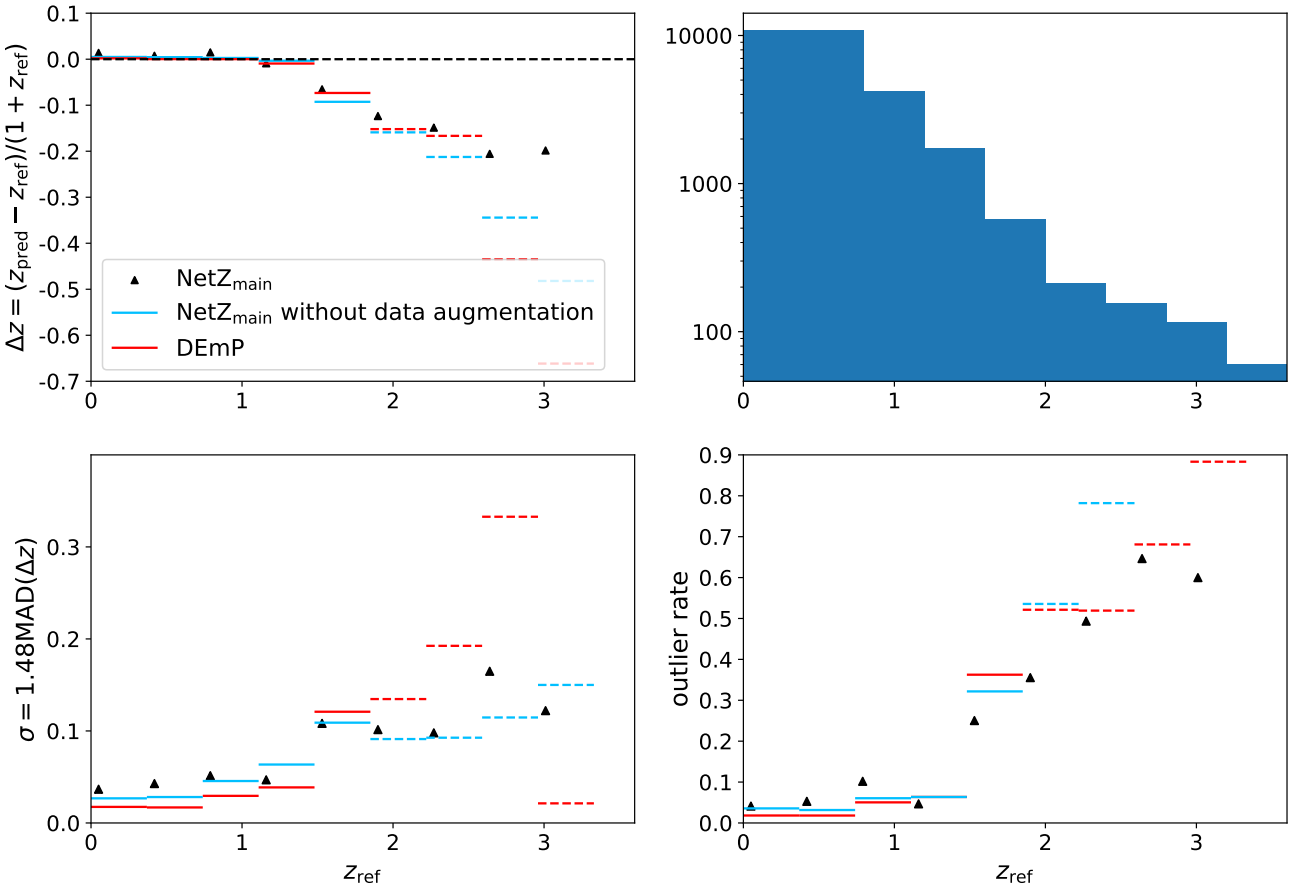

Fig. 4: NetZ$_{main}$ (black points) performance in terms of bias (top-left panel), dispersion (bottom-left panel), and outlier rate (bottom-right panel) as functions of the reference redshift z_{ref} in comparison to DEmP (red bars). Definitions of bias, dispersion and outlier rate are given in eq. 2-4. We show also with blue bars the results from a network where we do not use data augmentation to increase the number of high-z galaxies. Values in dashed bars are based on limited number of galaxies. The histogram in the top-right panel shows the number of galaxies as a function of redshift in the test set T used for the comparison. NetZ performs substantially better than DEmP at $z_{ref} \gtrsim 2$, with smaller bias, lower dispersion and lower outlier rate, by up to a factor of 2.

of NetZ$_{main}$ and with red bars the performance of DEmP. Since we use data augmentation (rotation and mirroring of images, see Section 2 for details) to increase the number of high-z galaxies which is not possible for DEmP, we trained also a network without data augmentation and show both in Figure 4 for comparison. For the range $z_{ref} \lesssim 1.5$, the performances of both methods are very good especially for the bias, with DEmP performing slightly better than NetZ. If we compare the range $z_{ref} \gtrsim 1.5$, the performance of both methods decreases, but NetZ with data augmentation does now noticeably better than DEmP. A decrease in performance in the redshift range around $z \approx 2$ is expected, as the used filter set *grizy* does not cover the prominent 4000Å break but, in contrast to the other methods presented in Tanaka et al. (2018), NetZ and DEmP can at least break the degeneracy between very low redshifts ($z < 0.2$) and the redshift range around 1.5 (i.e., the 4000Å break and Lyman break degeneracy). The very low dispersion of DEmP in the highest-redshift bin comes from DEmP underestimating consistently most of the redshifts, and hence the outlier rate is large. Although the outlier rate is high in the range $z_{ref} \gtrsim 2$ in general, the performance is primarily limited by the number of existing reference redshifts in this range. While DEmP is developed and tested with a big enough training sample also for higher redshifts, we used here for DEmP the exact same data set as for NetZ for a fair comparison. Since there is no sufficient training sample at $z_{ref} > 2$ for both NetZ without data augmentation and DEmP, we plot these values in dashed because they are less reliable. It is nonetheless encouraging to see the significant reduction in the bias, dispersion and outlier rate of NetZ with data augmentation for the high-z range, up to a factor of 2 relative to DEmP, by making use of the spatial information from the galaxy images in addition to photometry. Especially for upcoming surveys like LSST which will provide relatively deep images, it is important to have methods prepared and tested on the higher redshift range.

As a further comparison, we show in Figure 5 a scatter plot of z_{pred} versus z_{ref} for DEmP (Hsieh & Yee 2014) and our neural network NetZ$_{main}$ with data augmentation. From this plot we see again the good performance for the low-z range, where we note that the number of outliers from NetZ is negligible compared to the number of galaxies in the bins, which is also evident in the outlier rate. If we assume that all catastrophic outliers for $z_{ref} < 1.5$ are misplaced at high redshift, which is very conservative, then $> 77\%$ of the galaxies predicted to be at $z_{pred} > 1.5$ are actually at $z_{ref} > 1.5$. In the high-z range, NetZ tends to predict too low redshifts, but it does not have the cluster of catastrophic outliers at $z_{pred} \sim 0.5$ and z_{ref} between 3 and 3.5 that DEmP shows due to the Lyman-break/Balmer-break misclassifi-

cation. Even for the galaxies where the NetZ redshifts are classified as outliers, these redshifts are closer to the true redshift than for DEmP. The outlier rate for NetZ is dominated by blue-star-forming galaxies and galaxies with small spatial extent (covering $\approx 20-30$ pixels) that provide little information for the CNN to extract features. We therefore note that galaxies covering a small number of pixels are more prone to be catastrophic outliers in their redshifts and should be treated with caution.

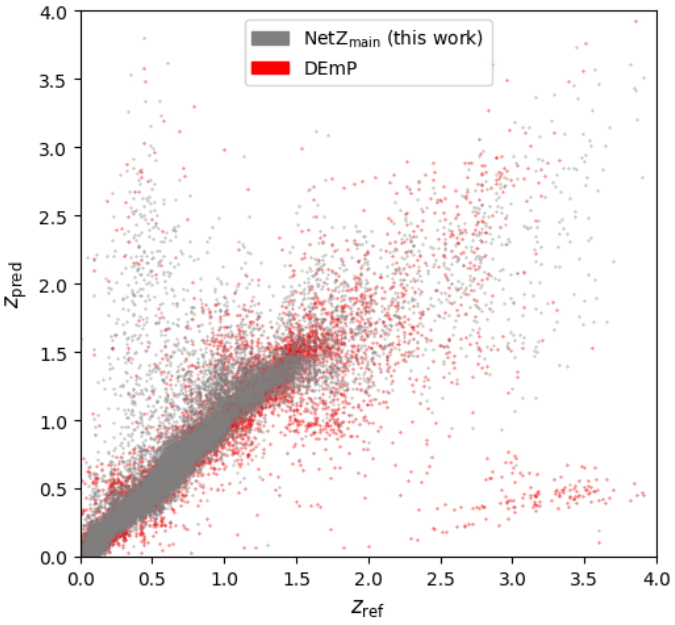

Fig. 5: Network performance as scatter plot comparing the predicted with the reference redshifts for NetZ$_{main}$ (this work) and DEmP (Hsieh & Yee 2014). The scatter looks overall comparable at $z_{ref} \lesssim 2$, while NetZ$_{main}$ does not contain the catastrophic outliers at $z_{ref} \sim 3$ and $z_{pred} \sim 0.5$ that DEmP has.

5.2. Photo-z with morphological information

The studies presented by, e.g., Soo et al. (2018) and Wilson et al. (2020) aim to include morphological information of galaxies to improve photo-z estimations. In particular, Wilson et al. (2020) make use of optical and near infrared observations, some of which are obtained with the Hubble Space Telescope (HST). Therefore the considered data cover a wider wavelength range and are additionally of better spatial resolution than our ground based HSC images. However, Wilson et al. (2020) consider only photometric measurements and 4 morphological measurements (half-light radius, concentration, asymmetry and smoothness), rather than the pixels directly. By working directly with the image pixels in our CNN, we use the maximal amount of information and are independent of the pipelines/uncertainties in extracting morphological measurements. Moreover, Wilson et al. (2020) limit the range to $0 < z < 2$, which makes the network not directly applicable to deep imaging surveys, especially as we focus on the high redshift range. As we show in the next section, we also obtain very good results in a limited range. With these differences in the assumptions and data sets, it is difficult to directly compare the results of Wilson et al. (2020) and NetZ. If we, nonetheless, compare our outlier fraction with ∼0.05 up to $z \sim 1.7$ (see Figure 4) to that from Wilson et al. (2020), which is called OLF, with ∼0.1-0.2 up to $z = 2$ (see Tables 2 and 3 of Wilson et al. (2020)), NetZ yields a good improvement. While Soo et al. (2018) and Wilson et al. (2020) find that morphological measurements do not provide a notable improvement in photo-z predictions when compared to using only multi-band photometric measurements, our NetZ results show that the pixels in the image cutouts that contain morphological information seem to be useful. This suggests that a promising avenue for future developments of photo-z methods is to combine photometric measurements (as typically used for current photo-z methods) with direct image cutouts (as used for NetZ) instead of morphological measurements.

5.3. Photo-z estimates for LSST

Schmidt et al. (2020) present a collection of different photo-z methods tested on LSST mock data. In detail, they compare 12 different codes, of which three methods are based on template fitting (BPZ, Benítez (2000); EAZY, Brammer et al. (2008); LePhare, Arnouts et al. (1999)), seven are based on machine learning (ANNz2, Sadeh et al. (2016); CMN,Graham et al. (2018); FlexZBoost, Izbicki et al. (2016); GPz, Almosallam et al. (2016a); METAPhoR, Cavuoti et al. (2017); SkyNet, Graff et al. (2014); TPZ, Carrasco Kind & Brunner (2013)), one is a hybrid method (Delight, Leistedt & Hogg 2017), and one is a pathological photo-z PDF estimator method (trainZ, Schmidt et al. 2020). The last method trainZ is designed to serve as an experimental control, and not a competitive photo-z PDF method. It assigns to each galaxy set a photo-z PDF by effectively performing a k-nearest neighbor procedure. As training data they use $< 10^7$ LSST like mock data limited to $0 < z < 2$ and an i band magnitude limit of 25.3 to match the LSST Gold Sample (for further details see Schmidt et al. 2020). The main advantage of these methods in Schmidt et al. (2020) compared to the current version of NetZ is the probability density function estimates, whereas NetZ does not require photo-z pre-selection and shows a good performance over a broader redshift range ($0 < z < 4$). Based on the different redshift range and data sets, a detailed and fair comparison is not possible. If we compare Figure 2 to Figure B1 of Schmidt et al. (2020) quantitatively, we see an overall similar performance, but most of the LSST methods have a cluster of outliers at $z_{ref} \sim 0.5$ and $z_{ref} \sim 1.7$ which we do not see with NetZ. The kink at $z_{ref} \sim 1.7$ might be related to the drop of data points and an edge effect near the end of the assumed range since we observe a similar effect with NetZ for higher redshifts ($z_{ref} \sim 3$). Comparing the machine learning methods is difficult as well. The network architectures, like nearest-neighbour algorithms, random forests, prediction trees or sparse Gaussian processes, presented in Schmidt et al. (2020) are simply too different from the image-based CNN we present with NetZ.

6. Limited-range and LRG-only redshift network

During our testing, we found substantial improvement by restricting the redshift range. We explored for instance networks with redshift ranges limited to $0 < z < 1$ and $1 < z < 2$, but not to higher redshift intervals due to limitations in available reference redshifts for $z > 2$. Limiting to $0 < z < 1$ is also done in several other publications (e.g., Hoyle 2016; Pasquet et al. 2019; Campagne 2020). To benefit from these refined networks in any practical application, one would first need to predict the correct redshift range and could then use these networks

in a specified range. One could also consider combining multiple networks and iteratively refining the photo-z predictions, i.e., start with NetZ$_{\rm main}$ to predict $z_{\rm pred}$, and based on the value of $z_{\rm pred}$, apply subsequently a network that is trained in a narrower range around $z_{\rm pred}$ to refine the $z_{\rm pred}$ estimate. However, we find that outliers from NetZ$_{\rm main}$ limit the gain we can achieve in refining $z_{\rm pred}$. A practical possibility to use a redshift network for the lower-redshift end of the distribution would be to restrict the sample by the galaxy brightness. If we restrict our data set D to galaxies with an apparent i-band AB magnitude brighter than 22, the catalog includes only 1.3% objects with $z_{\rm ref} \geq 1$ and we miss 12.9% of all galaxies from the original set D with $z_{\rm ref} \leq 1$. For the training of NetZ$_{\rm lowz}$ itself we limit only to a narrow redshift range but not in magnitude. The performance of NetZ$_{\rm lowz}$ is shown in Figure 6, on the left a histogram of the reference (red) and predicted (blue) redshifts. The distribution of the predicted redshift follows very well that of the reference redshift. On the right side we show a 1:1 correlation plot, with the median as a red line and in gray the 1σ and 2σ areas. If we compare to NetZ$_{\rm main}$ (Figure 2), we see that NetZ$_{\rm lowz}$ perform substantially better than NetZ$_{\rm main}$ as expected.

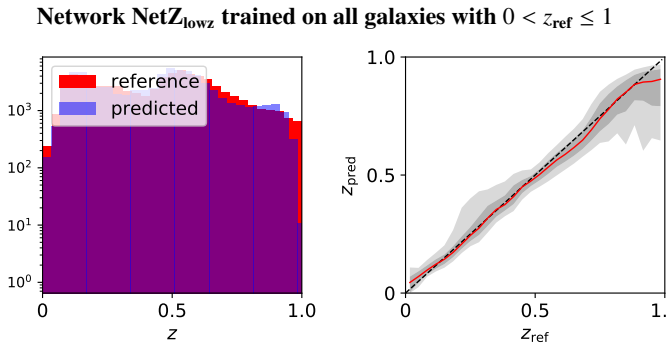

Fig. 6: Performance of the network NetZ$_{\rm lowz}$ trained on all types of galaxies in the range $0 < z_{\rm ref} \leq 1$. On the left hand side, histograms of the redshift distributions are shown, in red the distribution of the reference redshifts used to train the network (ground truth) and in blue the predicted redshift distribution. On the right panel a 1:1 comparison of reference and predicted redshifts is plotted. The red line show the median and the gray bands the 1σ and 2σ confidence levels.

We further show in Figure 7 the bias, dispersion and outlier rate for NetZ$_{\rm lowz}$ (red). If we compare this performance to NetZ$_{\rm main}$ applied to the same galaxies for a fair comparison (blue), we find a good improvement in the bias and, with a factor of ~ 2 reduction, in the dispersion. Only the outlier rate is comparable. If we compare to the network performance of NetZ$_{\rm main}$ on the full test set, we still see an improvement for the network NetZ$_{\rm lowz}$ without the i magnitude limitation. This confirms that the improvement is related to the network range. A scatter plot of NetZ$_{\rm lowz}$ is shown in Figure 8.

Instead of applying networks trained for specific redshift ranges, which is difficult to do in practice, we can consider specific classes of galaxies that can be selected a priori, such as Luminous Red Galaxies (LRGs). Therefore we investigate a redshift estimation network specialized on LRGs which are useful for various studies including strong lensing and baryon acoustic oscillations. Since nearly all LRGs out of our reference sample have $z_{\rm pred} < 1$, we show here the network performance of

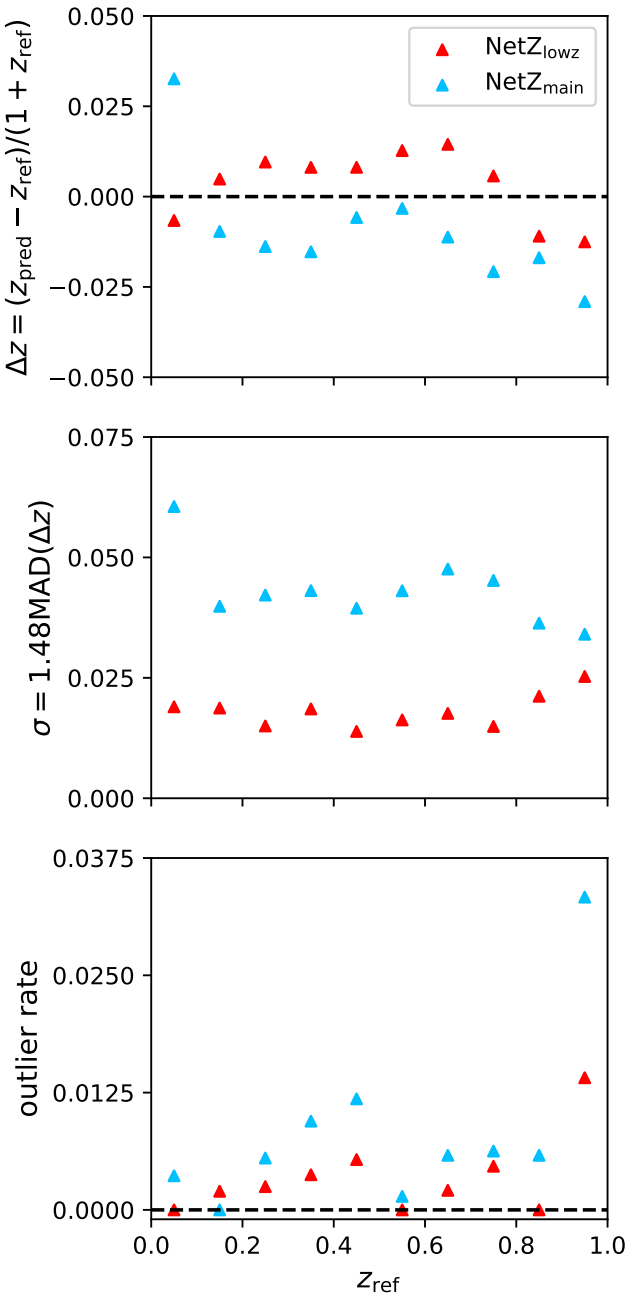

Fig. 7: Network performance of NetZ$_{\rm lowz}$ compared to NetZ$_{\rm main}$ in terms of bias, dispersion, and outlier rate (see eq. 2-4) as functions of the reference redshift $z_{\rm ref}$. For this comparison, we use the overlap between both test sets and only galaxies with an i-band magnitude brighter than 22 as NetZ$_{\rm lowz}$ would be applied only to them.

NetZ$_{\rm LRG}$ in comparison to the network NetZ$_{\rm lowz}$ trained on all galaxy types. Figure 9 shows on the left a histogram and on the right the 1:1 comparison of $z_{\rm ref}$ and $z_{\rm pred}$.

We show further the bias, dispersion, and outlier rate (defined in eq 2-4) in Figure 10. The network NetZ$_{\rm LRG}$ performs better in most redshift bins. Finally, in Figure 8 we show a scatter plot of this network without magnitude limitation in comparison to NetZ$_{\rm lowz}$. From this we can see again the redshift limits of the LRG sample and also the good improvement.

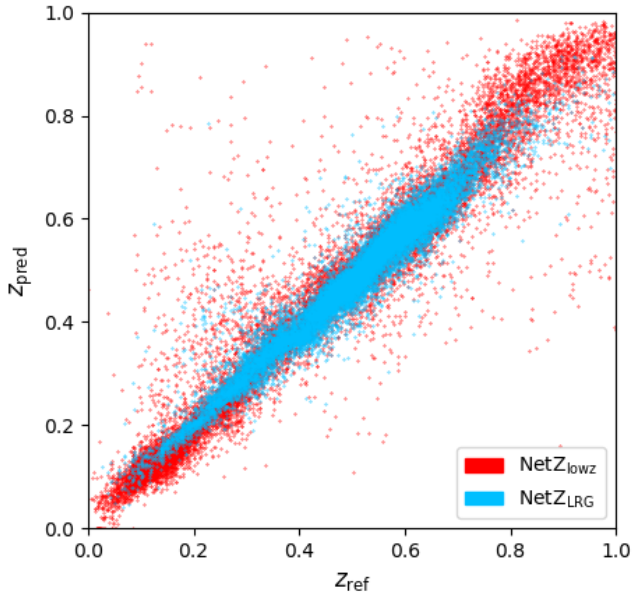

Fig. 8: Predicted redshifts $z_{\rm pred}$ against the reference redshifts $z_{\rm ref}$ for the networks NetZ$_{\rm LRG}$ and NetZ$_{\rm lowz}$ of their test set. We see directly a lower outlier rate for NetZ$_{\rm LRG}$ than NetZ$_{\rm lowz}$.

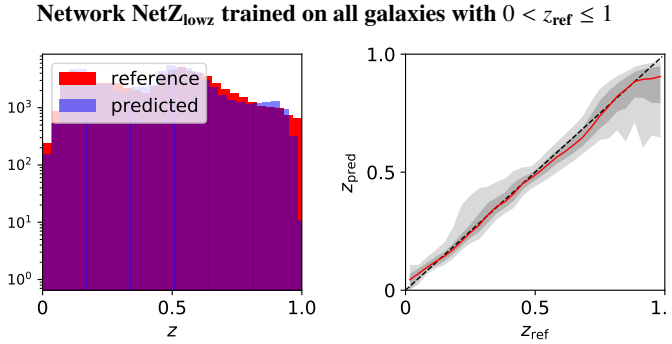

Fig. 9: Performance of the network NetZ$_{\rm LRG}$ (bottom) trained on LRGs only. On the left hand side, histograms of the redshift distributions are shown, in red the distribution of the reference redshifts used to train the network (ground truth) and in blue the predicted redshift distribution. On the right panel a 1:1 comparison of reference and predicted redshifts is plotted. The red line show the median and the gray bands the 1σ and 2σ confidence levels.

Both networks NetZ$_{\rm LRG}$ and NetZ$_{\rm lowz}$ show that photo-z for subsamples of galaxies does overall better than the main network NetZ$_{\rm main}$ that is trained on all galaxies. Therefore, for specific subsamples, it would be beneficial to train a CNN specific to that sample.

7. Summary and Conclusion

With the current and upcoming imaging surveys we expect billions of galaxies to be observed, while just a small fraction of them will have spectroscopically confirmed redshifts. Therefore, it is necessary to have tools to obtain good photometric redshifts, especially for the higher redshift range as the upcoming surveys will provide deeper images where previous photo-z methods have strong limitations. With the success of ML and especially CNNs in image processing, we investigated a new CNN based technique to estimate the photo-z of a galaxy. The method is very general; it accepts directly cutouts of the observed images and predicts the corresponding redshift. Therefore it is directly applicable to all HSC cutouts after applying simple cuts on the Kron radius and i band magnitude observables. For training the network and testing the performance we compare to reference redshifts from various, mostly magnitude-limited surveys. In this paper we focus on HSC data with a pixel size of 0.168″ and use the available five filters *grizy*.

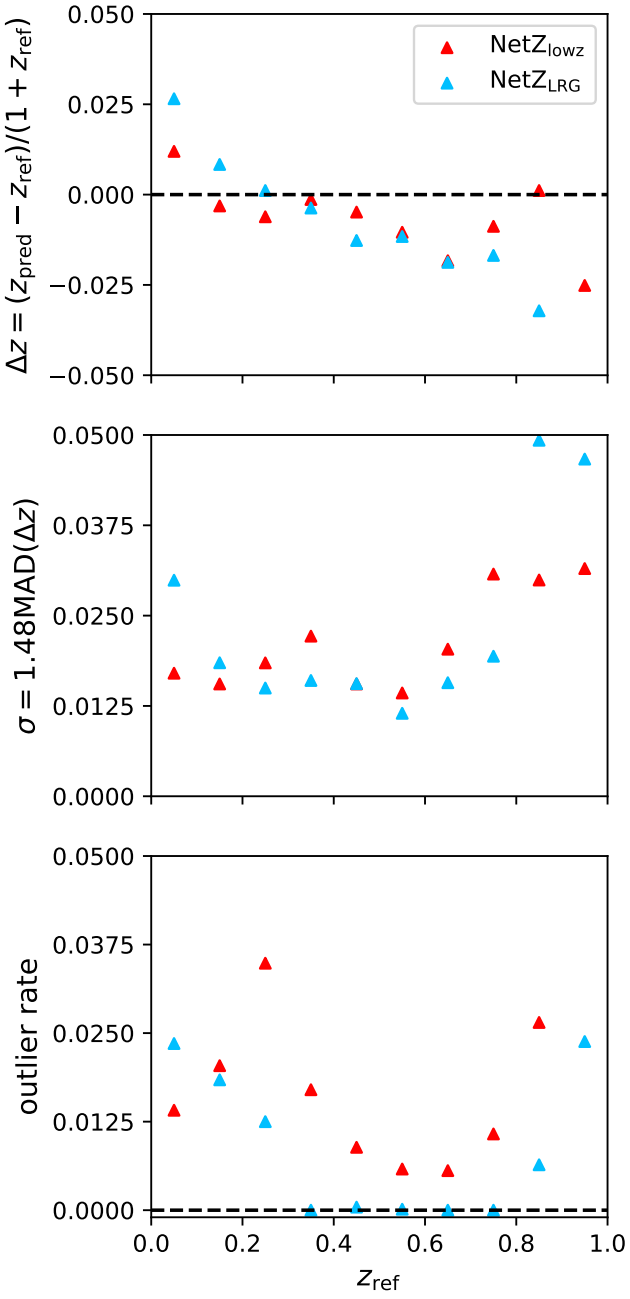

Fig. 10: Network performance of NetZ$_{\rm LRG}$ compared to NetZ$_{\rm lowz}$ applied to LRGs only in terms of bias, outlier rate, and dispersion (see eq. 2-4) as functions of the reference redshift $z_{\rm ref}$.

With our trained network on HSC images, we find an overall very good performance of the network with a 1σ uncertainty of 0.12 averaged over all galaxies from the whole redshift range. Our CNN provides a point estimate for each galaxy with uncertainties adopted from the scatter in each redshift bin of the test set. Based on the amount of available data, the network performs better in the redshift range below $z = 2$. In the range above $z = 2$, we are using – as an advantage over state-of-the art methods like DEmP – data augmentation by rotating and mirroring the images. While the bias for DEmP and NetZ$_{\rm main}$ as well as the dispersion for DEmP increases significantly in this range, with NetZ$_{\rm main}$ we obtain by using data augmentation similar values as for the lower redshift range. We obtain also better outlier rates for the highest redshift bins by using data augmentation but the improvement is less strong. In particular, NetZ does not underpredict the redshifts of galaxies with $z_{\rm ref} \sim 3-3.5$ by as much as DEmP and other methods due to the Lyman and Balmer break mis-classification. The main limitations that all photo-z methods face when predicting redshifts for distant galaxies is the low number of reference redshifts. In our case the number drops by a factor of around 1000 compared to the range $z < 2$. Therefore, using the image cutouts gives a good advantage as we can use data augmentation by rotating and mirroring the images. The effect is impressive as one can see in Fig 4. Since this is not possible for other photo-z methods, several of them are focusing only on the lower redshift range $z < 1$ or even lower (e.g., Hoyle 2016; Pasquet et al. 2019; Campagne 2020). If we also limit the redshift range to $z < 1$, we find a substantial improvement in our network's performance.

In case we focus on a specific galaxy type like LRGs, we find a further improvement of the network. This is understandable as the network can learn better the specific features of this galaxy type. Based on the small number of LRGs with redshift above $z = 1$, we limit the range of NetZ$_{\rm LRG}$ to $0 < z < 1$ and compare to a network trained on all galaxy types in the same redshift range for a fair comparison.

This paper provides a proof of concept to use a CNN for photo-z estimates. Based on the encouraging results of NetZ particularly at high redshifts, we propose further investigations along the lines of combining our CNN with a nearest-neighbor algorithm or a fully-connected network that ingests catalog-based photometric quantities (see Leal-Taixé et al. 2016). There are several methods, like DEmP and other methods (e.g. D'Isanto & Polsterer 2018; Schmidt et al. 2020), which provide a probability distribution function for the redshifts. Further developments of our CNN approach to provide a probability distribution function of the photo-z require more complex networks such as Bayesian neural networks (e.g., Perreault Levasseur et al. 2017) or mixture density networks (D'Isanto & Polsterer 2018; Eriksen et al. 2020). While this is beyond the scope of the current paper, such Bayesian or mixture networks are worth exploring.

In this paper, we have shown that a very simple convolutional neural network is able to predict accurate redshifts directly from the observed galaxy images. NetZ therefore has the advantage of using maximal information from the intensity pixels in the galaxy images, rather than relying on photometric/morphological measurements that could be prone to uncertainties/biases especially in images of blended galaxies. We have run NetZ$_{\rm main}$ on 34,414,686 galaxies from the HSC public data release 2 (PDR2) wide survey and provide the catalog here[4]. We flag all negative predictions and clear catastrophic outliers ($z_{\rm pred} > 5$), which are 15,043 and 3,314 objects, respectively, as -99. In Figure 11 we show a histogram of the newly available photo-z values (blue filled), whose distribution resembles the magnitude-limited sample of the cleaned COSMOS2015 (Laigle et al. 2018, orange histogram), which was scaled by a factor of 1010 to have the same sample size for direct comparison. This check shows that our NetZ predictions are producing a realistic galaxy redshift distribution expected for a depth similar to LSST.

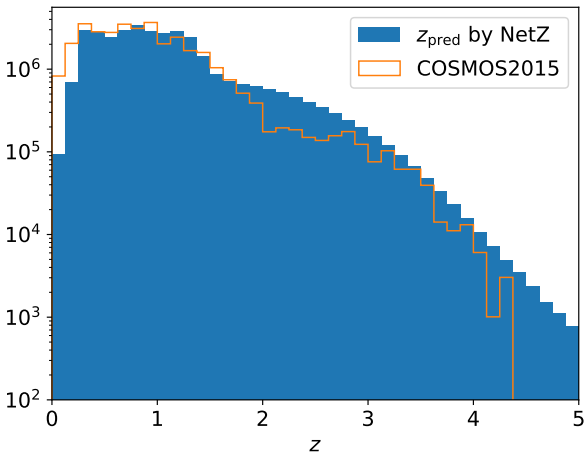

Fig. 11: Histogram of the newly predicted photo-z values with NetZ based on images of the HSC PDR2 (blue filled) and, for comparison, the distribution of COSMOS2015 (Laigle et al. 2018) scaled by a factor of 1010 to have the same sample size (orange open). The similarity in the two distributions shows that NetZ produces a realistic galaxy redshift distribution.

Since the image quality, depth and processing of HSC and LSST first-year data should be similar (the image processing pipeline of HSC is a branch of the LSST pipeline), the method we have developed here will be directly applicable and beneficial to LSST. The additional u-filter in LSST will likely further improve photo-z predictions. When applying our method to LSST, we do not expect the necessity to test the network architecture but probably some hyper-parameter combinations. Since the training is better done on real images rather than on mock images as done, e.g., by Schmidt et al. (2020, and references therein), we suggest to train a new network on LSST images as soon as they are available.

Acknowledgements. We thank Andreas Breitfeld from our IT group for helpful support.
SS, SHS, and RC thank the Max Planck Society for support through the Max Planck Research Group for SHS. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (LENSNOVA: grant agreement No 771776).
The Hyper Suprime-Cam (HSC) collaboration includes the astronomical communities of Japan and Taiwan, and Princeton University. The HSC instrumentation and software were developed by the National Astronomical Observatory of Japan (NAOJ), the Kavli Institute for the Physics and Mathematics of the Universe (Kavli IPMU), the University of Tokyo, the High Energy Accelerator Research Organization (KEK), the Academia Sinica Institute for Astronomy and Astrophysics in Taiwan (ASIAA), and Princeton University. Funding was contributed by the FIRST program from Japanese Cabinet Office, the Ministry of Education, Culture, Sports, Science and Technology (MEXT), the Japan Society for the Promotion of Science (JSPS), Japan Science and Technology Agency (JST), the Toray Science Foundation, NAOJ, Kavli IPMU, KEK, ASIAA, and Princeton University.

[4] The catalog is available under `https://www.dropbox.com/sh/grjfo0gkcxsj9n2/AAD-B7D6m7_1i6GGTX0Ionwja?dl=0`

This paper makes use of software developed for the Legacy Survey of Space and Time. We thank the LSST Project for making their code available as free software at http://dm.lsst.org
This paper is based in part on data collected at the Subaru Telescope and retrieved from the HSC data archive system, which is operated by Subaru Telescope and Astronomy Data Center (ADC) at National Astronomical Observatory of Japan. Data analysis was in part carried out with the cooperation of Center for Computational Astrophysics (CfCA), National Astronomical Observatory of Japan.
We make partly use of the data collected at the Subaru Telescope and retrieved from the HSC data archive system, which is operated by Subaru Telescope and Astronomy Data Center at National Astronomical Observatory of Japan.

References

Aihara, H., AlSayyad, Y., Ando, M., et al. 2019, PASJ, 106
Aihara, H., Arimoto, N., Armstrong, R., et al. 2018, PASJ, 70, S4
Alam, S., Albareti, F. D., Prieto, C. A., et al. 2015, The Astrophysical Journal Supplement Series, 219, 12
Almosallam, I. A., Jarvis, M. J., & Roberts, S. J. 2016a, MNRAS, 462, 726
Almosallam, I. A., Lindsay, S. N., Jarvis, M. J., & Roberts, S. J. 2016b, MNRAS, 455, 2387
Arnouts, S., Cristiani, S., Moscardini, L., et al. 1999, MNRAS, 310, 540
Benítez, N. 2000, ApJ, 536, 571
Bertin, E. & Arnouts, S. 1996, A&AS, 117, 393
Bolzonella, M., Miralles, J. M., & Pelló, R. 2000, A&A, 363, 476
Bonnett, C. 2015, MNRAS, 449, 1043
Bonnett, C., Troxel, M. A., Hartley, W., et al. 2016, Phys. Rev. D, 94, 042005
Bradshaw, E. J., Almaini, O., Hartley, W. G., et al. 2013, Monthly Notices of the Royal Astronomical Society, 433, 194–208
Brammer, G. B., van Dokkum, P. G., & Coppi, P. 2008, ApJ, 686, 1503
Campagne, J.-E. 2020, arXiv e-prints, arXiv:2002.10154
Carliles, S., Budavári, T., Heinis, S., Priebe, C., & Szalay, A. S. 2010, ApJ, 712, 511
Carrasco Kind, M. & Brunner, R. J. 2013, MNRAS, 432, 1483
Cavuoti, S., Amaro, V., Brescia, M., et al. 2017, MNRAS, 465, 1959
Coil, A. L., Blanton, M. R., Burles, S. M., et al. 2011, The Astrophysical Journal, 741, 8
Collister, A. A. & Lahav, O. 2004, PASP, 116, 345
Cool, R. J., Moustakas, J., Blanton, M. R., et al. 2013, The Astrophysical Journal, 767, 118
Coupon, J., Ilbert, O., Kilbinger, M., et al. 2009, A&A, 500, 981
Dahlen, T., Mobasher, B., Faber, S. M., et al. 2013, ApJ, 775, 93
Davis, M., Faber, S. M., Newman, J., et al. 2003, Discoveries and Research Prospects from 6- to 10-Meter-Class Telescopes II
D'Isanto, A. & Polsterer, K. L. 2018, A&A, 609, A111
Drinkwater, M. J., Jurek, R. J., Blake, C., et al. 2010, Monthly Notices of the Royal Astronomical Society, 401, 1429–1452
Eriksen, M., Alarcon, A., Cabayol, L., et al. 2020, arXiv e-prints, arXiv:2004.07979
Feldmann, R., Carollo, C. M., Porciani, C., et al. 2006, MNRAS, 372, 565
Garilli, B., Guzzo, L., Scodeggio, M., et al. 2014, Astronomy & Astrophysics, 562, A23
Graff, P., Feroz, F., Hobson, M. P., & Lasenby, A. 2014, MNRAS, 441, 1741
Graham, M. L., Connolly, A. J., Ivezić, Ž., et al. 2018, AJ, 155, 1
Hildebrandt, H., Arnouts, S., Capak, P., et al. 2010, A&A, 523, A31
Hildebrandt, H., Erben, T., Kuijken, K., et al. 2012, MNRAS, 421, 2355
Hildebrandt, H., Wolf, C., & Benítez, N. 2008, A&A, 480, 703
Hoyle, B. 2016, Astronomy and Computing, 16, 34
Hsieh, B. C. & Yee, H. K. C. 2014, ApJ, 792, 102
Izbicki, R., Lee, A. B., & Freeman, P. E. 2016, arXiv e-prints, arXiv:1604.01339
Laigle, C., Pichon, C., Arnouts, S., et al. 2018, MNRAS, 474, 5437
Le Fèvre, O., Cassata, P., Cucciati, O., et al. 2013, Astronomy & Astrophysics, 559, A14
Leal-Taixé, L., Canton Ferrer, C., & Schindler, K. 2016, arXiv e-prints, arXiv:1604.07866
Leistedt, B. & Hogg, D. W. 2017, ApJ, 838, 5
Lilly, S. J., Le Brun, V., Maier, C., et al. 2009, ApJS, 184, 218
Lima, M., Cunha, C. E., Oyaizu, H., et al. 2008, MNRAS, 390, 118
Liske, J., Baldry, I. K., Driver, S. P., et al. 2015, Monthly Notices of the Royal Astronomical Society, 452, 2087–2126
McLure, R. J., Pearce, H. J., Dunlop, J. S., et al. 2012, Monthly Notices of the Royal Astronomical Society, 428, 1088–1106
Momcheva, I. G., Brammer, G. B., van Dokkum, P. G., et al. 2016, The Astrophysical Journal Supplement Series, 225, 27
Newman, J. A., Cooper, M. C., Davis, M., et al. 2013, The Astrophysical Journal Supplement Series, 208, 5
Nishizawa, A. J., Hsieh, B.-C., Tanaka, M., & Takata, T. 2020, arXiv e-prints, arXiv:2003.01511
Pasquet, J., Bertin, E., Treyer, M., Arnouts, S., & Fouchez, D. 2019, A&A, 621, A26
Pasquet-Itam, J. & Pasquet, J. 2018, A&A, 611, A97
Perreault Levasseur, L., Hezaveh, Y. D., & Wechsler, R. H. 2017, ApJ, 850, L7
Rousseeuw, P. J. & Croux, C. 1993, Journal of the American Statistical Association, 88, 1273
Sadeh, I., Abdalla, F. B., & Lahav, O. 2016, PASP, 128, 104502
Schmidt, S. J., Malz, A. I., Soo, J. Y. H., et al. 2020, arXiv e-prints, arXiv:2001.03621
Silverman, J. D., Kashino, D., Sanders, D., et al. 2015, The Astrophysical Journal Supplement Series, 220, 12
Singal, J., Shmakova, M., Gerke, B., Griffith, R. L., & Lotz, J. 2011, PASP, 123, 615
Skelton, R. E., Whitaker, K. E., Momcheva, I. G., et al. 2014, The Astrophysical Journal Supplement Series, 214, 24
Soo, J. Y. H., Moraes, B., Joachimi, B., et al. 2018, MNRAS, 475, 3613
Tagliaferri, R., Longo, G., Andreon, S., et al. 2003, Neural Networks for Photometric Redshifts Evaluation, Vol. 2859, 226–234
Tanaka, M., Coupon, J., Hsieh, B.-C., et al. 2018, PASJ, 70, S9
Wilson, D., Nayyeri, H., Cooray, A., & Häußler, B. 2020, ApJ, 888, 83
Wolf, C. 2009, MNRAS, 397, 520