





# The domain adaptation problem in photometric redshift estimation: A solution applied to the HSC Survey

M. Treyer<sup>1,\*</sup>, R. Ait-Ouahmed<sup>1</sup>, S. Arnouts<sup>1</sup>, J. Pasquet<sup>2,3</sup>, E. Bertin<sup>4</sup>, G. Desprez<sup>5</sup>,  
V. Picouet<sup>6</sup>, and M. Sawicki<sup>7</sup>

<sup>1</sup> Aix Marseille Université, CNRS, CNES, LAM, Marseille, France

<sup>2</sup> AMIS, Université Montpellier Paul-Valéry, Montpellier, France

<sup>3</sup> TETIS – Inrae, AgroParisTech, Cirad, CNRS, Univ. Montpellier, Montpellier, France

<sup>4</sup> Université Paris-Saclay, Université Paris Cité, CEA, CNRS, AIM 91191, Gif-sur-Yvette, France

<sup>5</sup> Kapteyn Astronomical Institute, University of Groningen, PO Box 800, 9700AV Groningen, The Netherlands

<sup>6</sup> The California Institute of Technology, 1200 E. California Blvd., Pasadena, CA 91125, USA

<sup>7</sup> Department of Astronomy & Physics and Institute for Computational Astrophysics, Saint Mary's University, 923 Robie Street, Halifax, Nova Scotia B3H 3C3, Canada

Received 1 October 2025 / Accepted 8 December 2025

## ABSTRACT

**Context.** The multiband HSC-CLAUDS survey comprises several sky regions with varying observing conditions, only one of which, the COSMOS “Deep”, “Ultra Deep” and “Field” (UDF), offers extensive redshift coverage.

**Aims.** We aim to exploit a complete sample of labeled galaxies from the COSMOS UDF at  $i < 25$  ( $z \lesssim 5$ ) to train a convolutional neural network (CNN) and infer more accurate photometric redshifts in the other regions than those currently available from SED-fitting methods.

**Methods.** To address the severe domain mismatch problem that we observed when applying the trained CNN to regions other than the COSMOS UDF, we developed an unsupervised adversarial domain adaptation network that we grafted onto the CNN. The method was validated by three tests: the predicted redshifts were compared to the spectroscopic redshifts that are available for limited samples of mostly bright galaxies; the predicted redshift distributions of the entire galaxy population of a given field in several intervals of magnitude were compared to those of the COSMOS UDF, assumed to be representative; and the redshifts predicted for a sample of galaxies selected by narrow-band filter observations sensitive to [OII] emitters at  $z \sim 1.47$  were compared to those of confirmed [OII] emission line galaxies.

**Results.** The results show successful domain adaptation: the network is able to transfer its redshift classification capability learned from the COSMOS UDF to other regions of HSC-CLAUDS. Accuracy varies depending on magnitude and redshift, following that of the labels we used, but far exceeds that of currently available photometric redshifts. The catalogs of CNN redshifts we inferred for the XMM, DEEP2, and ELAIS fields and for the remaining COSMOS region ( $\sim 4$  million sources in total at  $i < 25$ ) are made public.

**Key words.** methods: data analysis – methods: numerical – galaxies: distances and redshifts

## 1. Introduction

Photometric redshifts are essential to cosmological surveys, since spectroscopy alone can no longer cover their extent. Template fitting techniques (e.g., Ilbert et al. 2006) have been in use for decades. They can predict redshifts in any interval, given an adequate set of spectral energy distribution (SED) templates and of photometric bands in which the galaxies are observed. However, for a given number of bands, they are outperformed by machine learning methods, provided that the training samples cover the full range of galaxy properties whose redshifts are to be predicted, which is not often the case. These methods are fast improving.

A multilayer perceptron (MLP) was first used to learn the mapping between photometry and redshift from data (ANNz by Collister & Lahav 2004; Vanzella et al. 2004). Other machine learning algorithms include Support Vector Machines (SVM; Wadadekar 2005), k Nearest Neighbours (kNN; Csabai et al. 2003; Zhang et al. 2013), Random Forest

(Carliles et al. 2010), XGBoost (Li et al. 2022), and Catboost (Li et al. 2024).

A significant step forward was made with the use of convolutional neural networks (CNNs), which can exploit the entire information content of multiband galaxy images to refine redshift estimates (Hoyle 2016; D’Isanto & Polsterer 2018). The benefits of CNNs were clearly demonstrated by Pasquet et al. (2019) (hereafter P19), who used a CNN trained with the SDSS five-band data (York et al. 2000) under a redshift bin classification loss. This proof-of-concept study illustrated how a CNN was able to make use of pixel-level information to extract features beyond colors and greatly improve redshift estimation compared to machine learning techniques. This work inspired a number of attempts to improve its results (e.g., Hayat et al. 2021; Dey et al. 2022; Treyer et al. 2024; Ait Ouahmed et al. 2024), some of them proposing hybrid models combining an MLP branch tasked with processing photometric features (e.g., magnitudes, colors), and a CNN branch that has access to multiband images of the galaxies (Menou 2019; Henghes et al. 2021, 2022; Yao et al. 2023; Zhang et al. 2024; Roster et al. 2024; Wei et al. 2025).

\* Corresponding author: marie.treyer@lam.fr

One condition for the applicability of CNN methods in a supervised learning context is the availability of a training sample that is sufficiently large and representative of the properties of the galaxy populations for which redshifts are to be inferred. Failure to comply with this condition leads to bias. Unfortunately it is difficult to meet with deep observations as they span large ranges of redshift and magnitude. Spectroscopy at faint magnitude is costly, and generally limited to small regions of large deep surveys, which causes yet another issue: CNNs fail to predict accurate redshifts in survey regions that differ observationally from those hosting the training sample, even if it is large and representative of the galaxies themselves.

In machine learning, this nuisance, known as the “domain adaptation problem,” occurs when the “target data” (on which the model will be used) does not come from the same distribution as the “source data” (on which the model is trained). The deep HSC-CLAUDS survey (Desprez et al. 2023), where labeled sources are essentially confined to one ultra deep field, provides us with an example of just such a problem, and an opportunity to try and remedy it.

Various deep domain adaptation (DA) approaches have been proposed to address this issue, including: (i) Discrepancy-based techniques, which minimize a divergence or distance between source and target feature distributions while jointly optimizing the main learning task (e.g., Long et al. 2015); (ii) Reconstruction-based DA (autoencoder- or VAE-based adaptation), in which an encoder shared between the source and target domains learns domain-invariant latent representations by optimizing a reconstruction objective in addition to the main task (e.g., Ghifary et al. 2015); (iii) Generative adversarial network (GAN)-based pixel-level DA, whereby GANs (Goodfellow et al. 2014) translate source images into target-like images to reduce the appearance gap. For example, PixelDA (Bousmalis et al. 2017) trains a generator to map source images into the target domain while preserving class labels; (iv) Adversarial feature-level DA, inspired by GANs but operating in feature space. These methods train feature extractors to be simultaneously predictive for the main task and invariant to the domain through adversarial domain classification (e.g., Ganin et al. 2016; Tzeng et al. 2017); (v) Transfer learning, commonly used in survey-to-survey or instrument-to-instrument adaptation. A model trained on a source dataset is partially retrained or fine-tuned on a small amount of labeled or unlabeled target data; (vi) Semi-supervised or weakly supervised DA, in cases where only limited or weak labels are available in the target domain; (vii) Universal or open-set DA, which address scenarios in which the target domain may contain new, missing, or unknown classes.

These DA methods, or combinations thereof, have been used in various astrophysical endeavors, mostly for classification tasks. Indeed, domain shifts are ubiquitous in astronomy: between simulated and observed data, between different instruments or surveys, between clean and noisy images, between different calibration systems, and so on. Discrepancy-based methods have been used, for example, to improve the robustness of galaxy morphology classification under degraded image quality (Ćiprijanović et al. 2022), to extract domain-independent cosmological information from different hydrodynamical cosmological simulations (Roncoli et al. 2023), and to align simulated and noisy images to estimate Einstein radii (in a regression task, Swierc et al. 2024). They have also been used in combination with adversarial DA between simulations and SDSS observations to classify galaxy mergers (Ćiprijanović et al. 2021). Density-ratio (importance-weighting) DA (a statistical discrepancy-based method) has been employed between

simulated and observed SEDs to improve the recovery of star-formation histories (Gilda et al. 2024). Reconstruction-based DA (VAE) has been used for cross-survey galaxy morphology classification (Xu et al. 2023). Adversarial feature-level DA has been applied between photo-ionisation models and integral-field observations to classify ionised nebulae (Belfiore et al. 2025), between HST and JWST images to transfer morphological classification (Huertas-Company et al. 2024), and between simulated lensing data and HSC data to find strong gravitational lenses (Alexander et al. 2023). Semi-supervised open-set DA has been used for cross-survey galaxy morphology classification enabling anomaly detection (e.g., mergers or strong lenses in the target domains) (Ćiprijanović et al. 2023). Transfer learning has been applied to adapt Galaxy Zoo morphology labels from the DECaLS survey to the BASS and MzLS components of the DESI Legacy Imaging Surveys using a fine-tuning approach (Ye et al. 2025), and to identify blended sources using networks pre-trained on natural images (Farrens et al. 2022).

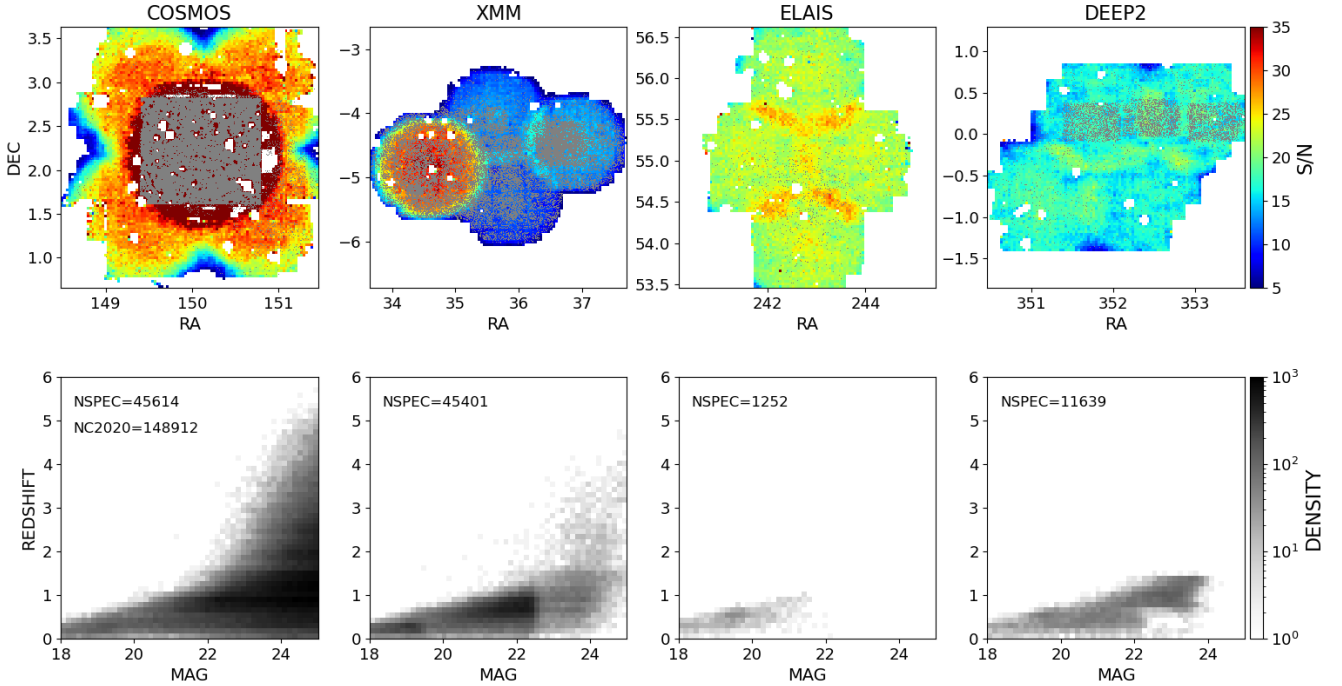
Here, we aim to implement DA to estimate CNN photometric redshifts across the entire HSC-CLAUDS survey from a unique labeled field. Discrepancy-based methods require a choice of distance metric that would effectively create domain-invariant representations and they are computationally expensive. Reconstruction-based methods may lose the redshift information in the reconstruction process as it is a subtle signal in the correlations between bands. The same goes for GAN-based methods, as it may be challenging to translate images from one field to another while preserving the redshift information. Transfer learning with fine tuning proved ineffective at faint magnitude given the shortage of labels in the target fields. Thus we chose to develop an adversarial approach, which would allow us to directly optimize a representation space that does not distinguish between fields and can be used for redshift estimation.

The paper is organized as follows: the HSC-CLAUDS data and the training set are described in Sections 2 and 3, respectively; the baseline redshift-estimating CNN and its DA issue are introduced in Section 4; and our adversarial solution and its application to the different HSC regions are presented in Sections 5 and 6, respectively. We highlight shortcomings in Section 7 and conclude in Section 8. Samples of images and results are displayed in Appendix A.

## 2. The HSC data

### 2.1. The photometric data

The Hyper Suprime-Cam Strategic Survey Program on the Subaru telescope (HSC-SSP, Aihara et al. 2018) represents the current state of the art in deep, wide-area imaging surveys. The HSC Deep component is an imaging survey in the *grizy* filters covering 26 deg<sup>2</sup> to a limiting magnitude of  $i_{AB} \sim 26$ . This is an unprecedented combination of area and depth that will remain unmatched for many years after LSST scientific operations begin. It consists of four regions: COSMOS, XMM-LSS, ELAIS-N1, and DEEP2-3, each mapped by several pointings. We refer to these regions as simply COSMOS, XMM, ELAIS, and DEEP2. An ultra deep layer, 1 magnitude deeper, completes the survey in two small subregions (covering  $\sim 3.5$  deg<sup>2</sup>) of the COSMOS and XMM fields. We refer to these as the COSMOS and XMM ultra deep fields (UDFs), and to the non-UDF parts as COSMOS DEEP and XMM DEEP. We use the public DR2 (Aihara et al. 2019) for the Deep and Ultra Deep layers of the survey. These have median depths of  $g = 26.5$ ,  $y = 24.5$ , and  $g = 27$ ,  $y = 25.5$ , respectively.



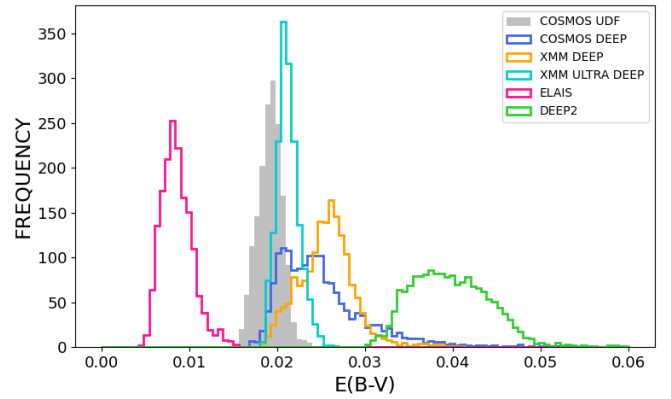
**Fig. 1.** *Top:* the four HSC regions color-coded by the  $i$ -band S/N of sources at  $24 < i < 25$ . Only sources imaged in all six bands are shown. Several masks were applied to rid the catalogs of stellar sources. The dark red circles in COSMOS and XMM are the two UDFs. The gray dots mark the position of sources with known redshifts at  $i < 25$ . *Bottom:* their distributions in the  $i$ -band magnitude (MAG)-redshift space. NSPEC is the number of galaxies with spectroscopic redshifts. NC2020 is the number of galaxies with redshifts from the COSMOS2020 photometric redshift catalog (introduced in Section 2.2).

Follow-up observations in the  $u$  band and in the slightly redder  $u^*$  band from the deep CLAUDS survey using the CFHT MegaCam imager (Sawicki et al. 2019) were added to this dataset. COSMOS was observed with both filters, while XMM was only covered by the filter  $u^*$ , and ELAIS and DEEP2 by the filter  $u$ . CLAUDS covers  $18 \text{ deg}^2$  of the four fields to a median depth  $u = 27$ , and  $1.6 \text{ deg}^2$  of the two UDFs to  $u = 27.4$ .

Deep near-infrared (NIR) observations in three bands ( $JHK_s$ ) are also partially available on the COSMOS field (McCracken et al. 2012) from the UltraVISTA<sup>1</sup> survey, and on XMM (Jarvis et al. 2013) from the VIDEO<sup>2</sup> survey. We did not use them in the present analysis due to their limited coverage, although we tested that adding them noticeably improved the CNN performance in the COSMOS field (training is described in Section 4.2), particularly at faint magnitude and high redshift.

The top panels of Fig. 1 show the four HSC regions color-coded by the  $i$ -band signal-to-noise ratio (S/N) of sources at  $24 < i < 25$ . The two UDFs are well identified as dark red circles (highest S/N) within COSMOS and XMM. There are also variations in seeing and observing conditions in the other filters. The spatial distributions of sources with known redshifts are overlaid. Their distributions in the magnitude-redshift space are displayed in the bottom panels and discussed in the next section. The E(B-V) distributions in the various subregions are shown in Fig. 2.

For each source at  $i < 25$ , we created a datacube of  $6 \times 64 \times 64$  pixels that includes one CLAUDS image ( $u^*$  or  $u$ ) and five HSC images (*grizy*). All images were projected onto the same HSC reference pixel grid, using SWARP



**Fig. 2.** Normalized E(B-V) distributions in the different HSC subregions.

(Bertin et al. 2002), with a scale of  $0.168 \text{ arcsec/pixel}$ . Sources whose images presented obvious defects were excluded. Most stellar and dubious sources were also removed with the following mask:  $\text{CLASS\_STAR\_HSC\_I} < 0.9$  &  $\text{COMPACT} == 0$  &  $\text{OBJ\_TYPE} == 0$  &  $\text{MASK} == 0$ . We chose to limit our analysis to  $i = 25$  because the accuracy of our redshift estimation is too severely degraded beyond this magnitude, as shown in following sections.

SED-fitting redshifts were computed for all sources in all four regions, based on *ugrizy* magnitudes or on *ugrizy* +  $JHK_s$  magnitudes when available in the COSMOS and XMM fields. The purpose of this work is to supersede them with estimates of significantly higher quality. A full description of the HSC Deep survey and its ancillary data can be found in Desprez et al. (2023) and Picouet et al. (2023).

<sup>1</sup> <https://ultravista.org>

<sup>2</sup> <https://www.eso.org/sci/publications/announcements/sciann17491.html>

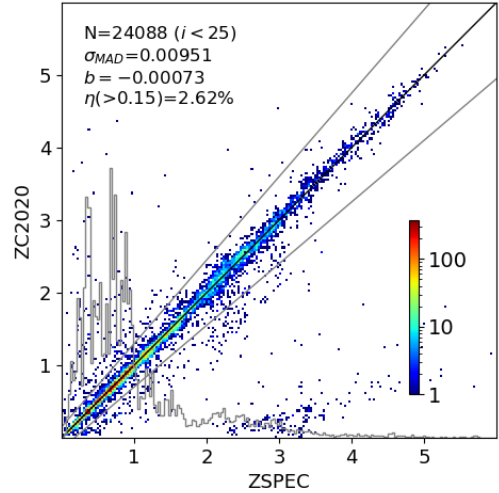
**Table 1.** Summary of the spectroscopic surveys with their typical spectral resolution, redshift range, and main target selection criteria.

Spectroscopy			
Survey	Res.	$z$ -range	Selection
SDSS DR12 <sup>(1)</sup>	2000	$z \leq 0.4$	$r \leq 17.8$
SDSS-BOSS <sup>(2)</sup>	2000	$0.3 \leq z \leq 0.7$	LRGs
GAMA <sup>(3)</sup>	1300	$z \leq 0.7$	$r \leq 19.8$
WIGGLEZ <sup>(4)</sup>	1300	$z \leq 1.2$	$NUV \leq 22.8$
zCOSMOS <sup>(5)</sup>	650	$z \leq 1.2 - 5$	$r \leq 22.5 - 25$
VANDELS <sup>(6)</sup>	650	$1 \leq z \leq 6$	$H \leq 25$
UDSz <sup>(7)</sup>	650	$z \leq 4$	$K \leq 23$
DEEP2 <sup>(8)</sup>	6000	$0.7 \leq z \leq 1.5$	$r \leq 24$
VVDS <sup>(9)</sup>	230	$z \leq 1.2 - 6$	$i \leq 22.5 - 24$
VIPERS <sup>(10)</sup>	230	$0.4 \leq z \leq 1.5$	$i \leq 22.5$
VUDS <sup>(11)</sup>	230	$2 \leq z \leq 6$	$K \leq 23$
CLAMATO <sup>(12)</sup>	1100	$2 \leq z \leq 3.5$	LBGs
C3R2 <sup>(13)</sup>	1100	$z \leq 4$	SOM
COSMOS <sup>(14)</sup>	Multiple	$z \leq 4$	multiple
3DHST <sup>(15)</sup>	130	$z \leq 4$	$H \leq 24$
PRIMUS <sup>(16)</sup>	40	$z \leq 0.9$	$i \leq 22.5$
COSMOS20 <sup>(17)</sup>	Photo- $z$	$z \leq 6$	$i \leq 26.5$

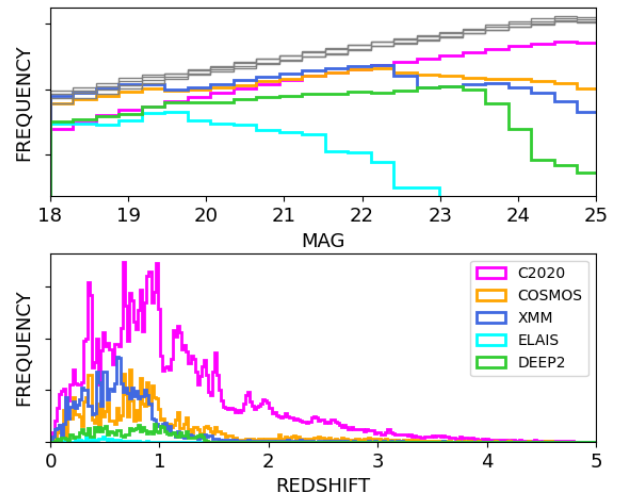
## 2.2. The labeled data

The spectroscopic redshifts available in the HSC fields come from a compilation of spectroscopic surveys summarized in Table 1. The COSMOS field is completed with redshifts from the Khostovan et al. (2025) compilation. A total of  $\sim 104\,000$  secure spectroscopic redshifts are available at  $i < 25$ , unevenly distributed among the four regions and in magnitude-redshift space (84% are in COSMOS and XMM), as can be seen in the lower panels of Fig. 1. In the COSMOS UDF, the lack of completeness in magnitude-redshift space is compensated for by the COSMOS2020 photometric redshift catalog (Weaver et al. 2022). These redshifts were computed from 30 photometric bands, ranging from UV to IR, via SED-fitting techniques in the UDF. The authors estimated four photometric redshifts based on two different multiband photometric catalogs (using two distinct flux extraction software packages) and two different SED-fitting codes. To construct a reliable sample, we computed the mean and standard deviation of these four redshifts,  $\bar{z}$  and  $\sigma_z$ , and retained those with  $\sigma_z \leq 0.1(1 + \bar{z})$ . Figure 3 illustrates the quality of this selection (hereafter the C2020 sample) for sources with measured spectroscopic redshifts<sup>3</sup>. The global deviation,  $\sigma_{MAD}$ , and median bias,  $b$ , (see Section 4.2) quoted in the figure, are remarkably low. The fraction of catastrophic outliers,  $\eta$ , could be improved by restricting the sample to the most secure spectroscopic redshifts but we chose to keep them all as our labels are imperfect in any case. Indeed, the spectroscopic sample is essentially confined to  $i < 22.5$  and  $z < 2$  and the C2020 metrics degrade with magnitude and redshift.

The  $i$ -band magnitude and redshift distributions of the spectroscopic samples per region and of the C2020 sample are



**Fig. 3.** C2020 photometric redshifts (ZC2020) compared to spectroscopic redshifts (ZSPEC) when available in the COSMOS UDF at  $i < 25$ . Their distribution is overlaid in gray. The number of sources,  $N$ , and the C2020 performance metrics (Section 4.2) are quoted in the upper left corner.



**Fig. 4.** *Top:*  $i$ -band magnitude distributions of the photometric (gray), C2020 (pink), and spectroscopic samples per region in log scale. *Bottom:* C2020 and spectroscopic redshift distributions per region at  $i < 25$ .

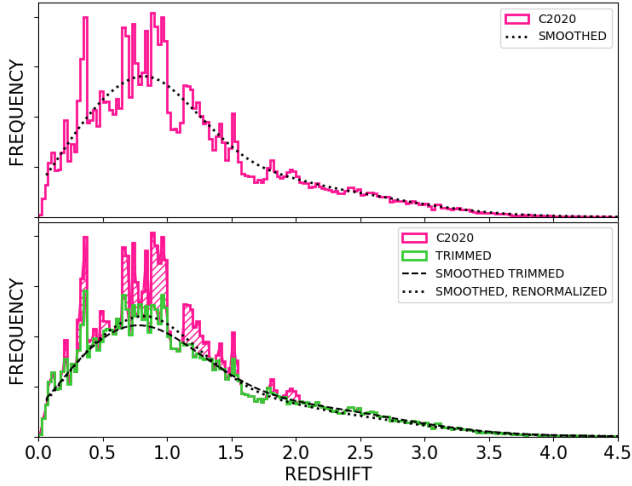
displayed in Fig. 4. The C2020 sample is complete to  $i = 25$ , as shown in the top panel where its magnitude distribution (in pink) can be compared to those of the complete photometric samples in the four regions (gray lines).

Finally, deep observations in the narrow-band filter NB921, picking out strong [OII] emitters at  $z \sim 1.47$  (Hayashi et al. 2020), complete the spectroscopic data. Although they may include a fraction of other galaxy types at other redshifts, they provide an additional test population.

## 3. The training set

Short of representative spectroscopy, the C2020 photometric redshift sample is currently the best training set at hand. To make the most of our resources, we replaced these redshifts by spectroscopic ones where available, essentially at low redshift and bright magnitudes ( $\sim 19\%$  of the C2020 sample at  $i < 25$ ), and added  $\sim 2700$  spectroscopic redshifts of bright COSMOS

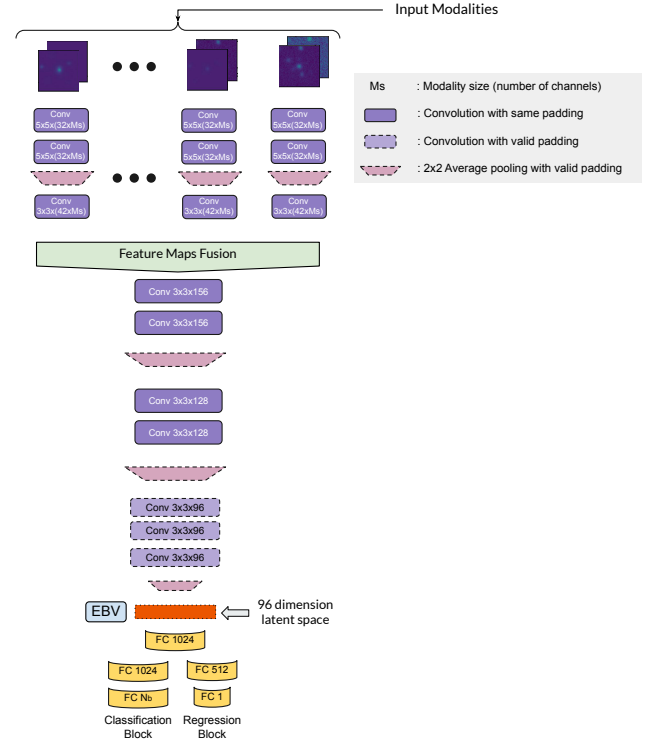
<sup>3</sup> While we use  $z$  as the usual notation for redshift in the text, with specifiers when necessary, e.g.,  $z_{\text{SPEC}}$ , in all figures we prefer to use capital letters for greater visibility: thus ZSPEC stands for spectroscopic redshift ( $z_{\text{SPEC}}$ ), ZC2020 for C2020 redshift ( $z_{\text{C2020}}$ ), ZCNN for CNN predicted redshift ( $z_{\text{CNN}}$ ), and ZPHOT for  $ugrizy$  or  $ugrizy+NIR$  SED-fitting photometric redshift ( $z_{\text{PHOT}}$ ). MAG stands for  $i$ -band magnitude.



**Fig. 5.** *Top panel:* simple smoothing of the training redshift distribution using SciPy’s KDE function. *Bottom panel:* galaxies from the highest-density regions in each of the large spikes protruding above the smoothed distribution were removed from the initial sample, resulting in the green histogram, which was then smoothed using KDE (dashed line). The hashed pink spikes show the number and redshift location of the discarded galaxies ( $\sim 14\%$  in total at  $i < 25$ ).

UDF galaxies without C2020 measurements. All the redshifts in this sample of  $\sim 179\,600$  galaxies, whether spectroscopic or C2020, were used as one-hot labels (assigned to a unique class of our classification scheme, described in Section 4). As the COSMOS field was mapped with both the  $u$ -band and  $u^*$ -band filters, whereas the other fields were mapped with only one of them, we created two sets of training data cubes: one containing  $ugrizy$  images and one containing  $u^*grizy$  images ( $\sim 179\,600 \times 6 \times 64 \times 64$  pixels each).

As a narrow pencil beam survey, the COSMOS UDF intercepts several structures resulting in large redshift spikes clearly visible in the bottom panel of Fig. 4 (the pink histogram). These redshift over-densities tend to bias the photometric redshift estimates of trained algorithms. They are also likely to encompass dense galaxy clusters containing an excess of massive red galaxies relative to the general population, potentially inducing an additional bias by increasing the probability of classifying such galaxies in other fields at these particular redshifts. We therefore chose to selectively smooth the data as follows. We first smoothed the original training redshift distribution in several bins of magnitude using the kernel density estimation (KDE) function from the SciPy library (Virtanen et al. 2020). This is shown as the dotted line in the upper panel of Fig. 5 for the full sample at  $i < 25$ . In the redshift intervals protruding above the smoothed distribution, we randomly removed galaxies from the highest-density regions identified in RA-Dec space, and then smoothed the trimmed distribution using KDE. The result is illustrated in the lower panel of Fig. 5. The hashed pink spikes show the number and redshift location of the galaxies in spatial over-densities that were excluded from the original distribution ( $\sim 14\%$  in total at  $i < 25$ ). The dashed line shows the final smoothed distribution, while the dotted line shows the same smoothed distribution as in the upper panel, renormalized to match the number of galaxies in the final one:  $\sim 155\,000$ . The difference is mild, as is the difference between the redshifts resulting from the first and second smoothed training distributions. However it is systematically in favor of the second step when appraisable. In the following, the training set refers to



**Fig. 6.** Baseline model (Ait Ouahmed et al. 2024): several modalities first process subsets of the input images consisting of two adjacent bands (five in our case). Their respective outputs are then merged and fed into a common series of convolutions.

this trimmed and smoothed dataset of  $\sim 155\,000$  galaxies from the COSMOS UDF, with C2020 redshifts for the most part, or spectroscopic redshifts. We expect this training set to be representative of the redshift distribution of galaxies at  $i < 25$  in all fields. As the process involves random draws from the full sample, it varies slightly at the start of each training epoch.

## 4. The baseline model

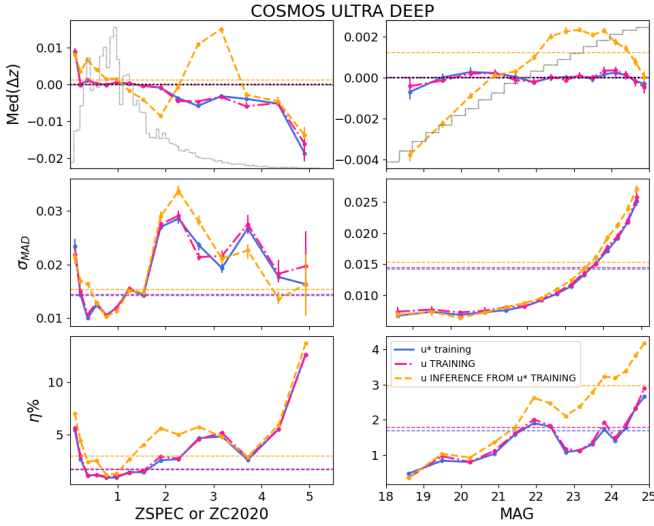
### 4.1. Baseline network

Our baseline model, shown in Fig. 6, is the multimodal CNN introduced by Ait Ouahmed et al. (2024), which improved on P19. It features five multimodalities upstream, each processing pairs of input images from two adjacent photometric bands. Their respective features maps are merged and fed into the main network. These multimodalities added to the main CNN were shown to improve redshift estimation. The network was trained as a classifier into 360 contiguous redshift bins between  $z = 0$  and 6. We considered the normalized output of the classifier to be a probability density function (PDF) and used its median value as the optimal redshift point estimate (Treyer et al. 2024).

### 4.2. Training

The following metrics were used to quantify the performance of redshift predictions ( $z_{\text{CNN}}$ ,  $z_{\text{C2020}}$ , and  $z_{\text{PHOT}}$ ):

- the **normalized residuals**,  $\Delta z = (z_{\text{PRED}} - z_{\text{TRUE}})/(1 + z_{\text{TRUE}})$ ,
- the **prediction bias**,  $b = \text{Med}(\Delta z)$  (median of the residuals),
- the **deviation**,  $\sigma_{\text{MAD}} = 1.4826 \times \text{MAD}$ , where MAD (median absolute deviation) =  $\text{Med}(|\Delta z - \text{Med}(\Delta z)|)$ ,
- the **fraction,  $\eta$ , of outliers** with  $|\Delta z| > 0.15$ ,



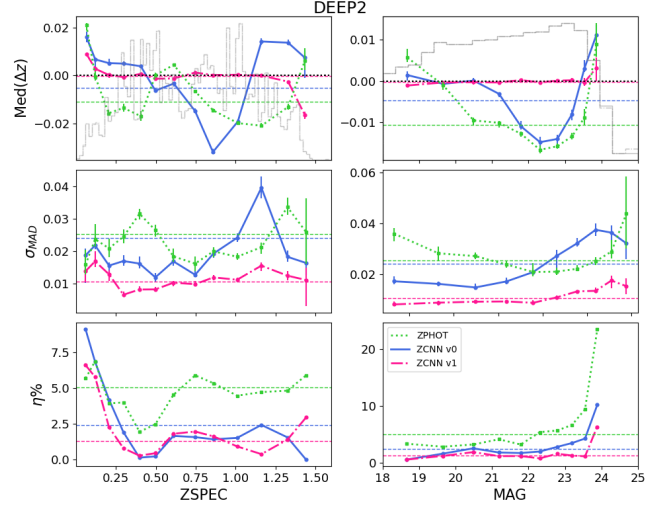
**Fig. 7.** CNN metrics as a function of redshift label and  $i$ -band magnitude for the  $u^*$  *grizy* and *ugrizy* validation sets, and for the  $u$  dataset inferred from the  $u^*$ -trained model. Using either set of bands for training makes no difference but inferring one from the other significantly degrades the performance. In this and all similar figures, the redshift and magnitude distributions are overlaid in gray in the corresponding upper panels, and the  $1\sigma$  error bars are based on 100 bootstrap samples.

where  $z_{\text{PRED}}$  and  $z_{\text{TRUE}}$  refer to the predicted and “true” redshifts (the labels), respectively. We first trained the network on both training datasets (*ugrizy* and  $u^*$  *grizy*) with a cross-validation protocol (80% training and 20% validation). The PDF predictions of each validation fold were averaged over six runs. Figure 7 shows the prediction bias, MAD and outlier fraction as a function of the redshift label (spectroscopic or C2020) and  $i$ -band magnitude for the  $u^*$  *grizy* model, the *ugrizy* model, and the  $u^*$  *grizy*-trained model applied to the *ugrizy* data. The set of bands that is used makes no difference but inferring one from the other significantly degrades the performance. Therefore we used whichever band was suited for each of the other fields ( $u^*$  for XMM and COSMOS DEEP,  $u$  for ELAIS and DEEP2). As was expected from the declining quality of the images (P19), the metrics significantly degrade with magnitude, hence our choice to limit the present analysis to  $i < 25$ . They also do at  $z < 0.5$ , probably due to the paucity of low-redshift galaxies in the training set and to the C2020 outliers seen in Fig. 3.

#### 4.3. Inference

From the above baseline model, we inferred baseline redshifts, referred to as v0 redshifts, for the spectroscopic (labeled) population ( $\sim 11\,600$  galaxies) and for the photometric (unlabeled) population ( $\sim 714\,000$  galaxies) in the DEEP2 field, as an illustration. In a second experiment, we enriched the COSMOS UDF training sample with the spectroscopic galaxies from the DEEP2 field and retrained the network with a cross-validation protocol (20% validation, 80% training, six ensembles). We refer to this model as v1.

Figure 8 shows the v0 and v1 CNN metrics for the DEEP2 spectroscopic sample as a function of redshift and magnitude, compared to the *ugrizy* SED-fitting redshifts. Except for the fraction of outliers, the v0 model is not a significant improvement on those. The deviation is even worse at  $22 < i < 24$ . However the gain is unambiguous for all three metrics with the v1 model.



**Fig. 8.** v0 and v1 CNN metrics (blue and pink lines, respectively. See Section 4.3) for the DEEP2 spectroscopic sample as a function of redshift and magnitude, compared to the *ugrizy* photometric redshifts (green lines).

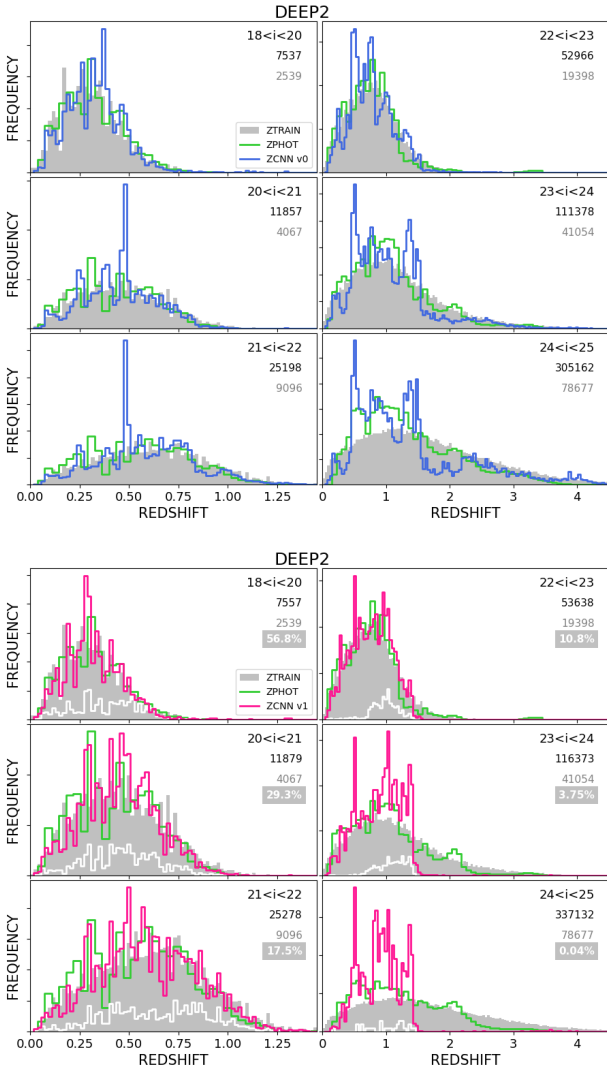
Figure 9 shows the redshift distributions of the complete, mostly unlabeled, DEEP2 sample in six intervals of magnitude inferred from the v0 and v1 models (top and bottom panels respectively). The distributions are normalized and compared the SED-fitting redshift distributions and to the training distributions, which are assumed to be representative and whose numbers of galaxies per magnitude bin are quoted in gray. All the predicted distributions are roughly consistent with the training distributions at  $i < 22$ . The agreement degrades at fainter magnitudes where the distributions from both CNN models become significantly, but differently, distorted.

The v1 redshift distributions are more inadequate than the v0 predictions. The network has classified all DEEP2 sources within the redshift range of the few DEEP2 spectroscopic sources included in the training set (white histograms), however small their contribution (quoted in white on gray). This suggests that it has primarily learned to distinguish between DEEP2 and COSMOS images and that it classifies the redshifts of DEEP2 sources based on DEEP2 images and labels present in the training set rather than on redshift-related patterns that could be learned from the much larger number of COSMOS images. The improvement seen in Fig. 8 is therefore very misleading: the network has learned to estimate the redshift of DEEP2 spectroscopic sources from other DEEP2 spectroscopic sources drawn from the exact same distribution, and to ignore COSMOS sources as irrelevant to the task.

DEEP2 illustrates what is known as the “domain adaptation problem”, the inability of the network to use information from one field to predict adequate results in another. This problem occurs, more or less, in all the other HSC regions (Section 6).

## 5. The adversarial model

The COSMOS UDF differs from the other fields in a number of ways, including S/N (Fig. 1), Galactic dust extinction (Fig. 2), and seeing. Whatever recipe the network learns from the COSMOS UDF images to estimate redshifts fails to work adequately with images from other sky regions. To try and blind the CNN to these differences, we developed the adversarial approach described below.

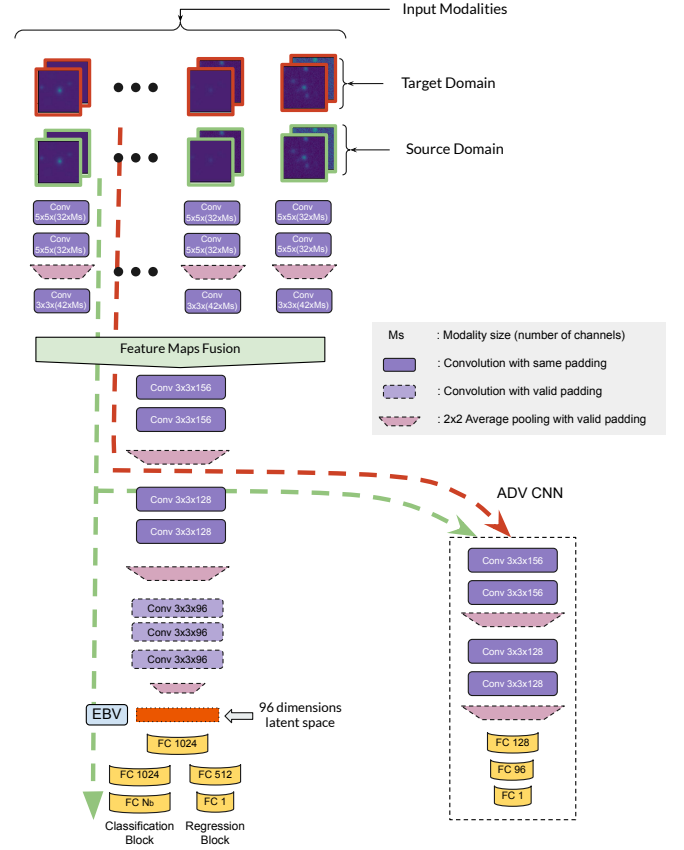


**Fig. 9.** Redshift distributions of the complete DEEP2 data inferred from the baseline CNN trained on the COSMOS UDF sample alone (v0, blue histograms) and on the COSMOS UDF data supplemented by the DEEP2 spectroscopic data (v1, pink histograms). The distributions are compared to the *ugrizy* SED-fitting redshift distributions (green histograms) and to the COSMOS UDF training distributions (shaded gray histograms). All are normalized to 1. The white histograms are the redshift distributions of the DEEP2 spectroscopic sources included in the training set, arbitrarily scaled for visibility. The numbers of DEEP2 galaxies and COSMOS UDF training galaxies and the fraction of DEEP2 training galaxies with respect to the latter are quoted in black, gray, and white on gray, respectively, in each magnitude range.

### 5.1. Architecture and training procedure

Figure 10 shows how we integrated an adversarial module into the CNN architecture depicted in Fig. 6. This module is composed of four convolutional layers: two with 156 kernels and two with 128 kernels, all of size  $3 \times 3$ . Every pair of convolutional layers is followed by a pooling layer. The module ends with fully connected layers generating the adversarial classification.

The network was trained with the source domain (the COSMOS UDF training set) and a target domain (100 000 sources randomly drawn from the photometric datacube of a field other than the COSMOS UDF. No labels are required in the target domain). At each iteration, a mini-batch was sampled from each domain. The source domain batch was used both for the redshift



**Fig. 10.** Domain adaptation architecture with the added adversarial module.

classification training and for the adversarial objective, while the target batch only served the adversarial objective.

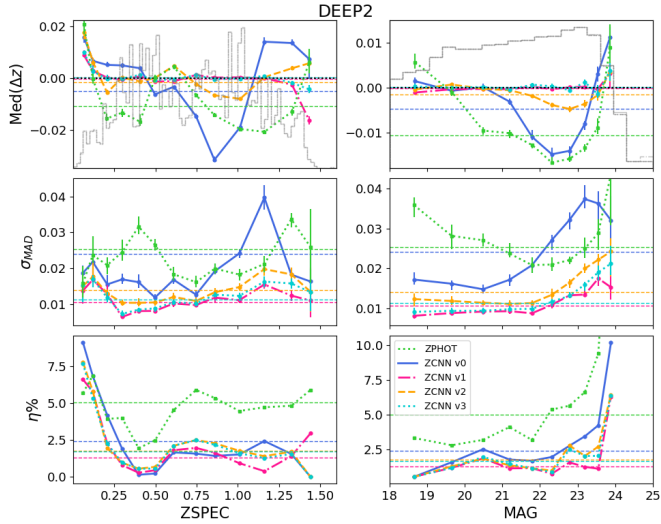
The adversarial module receives feature maps from a layer of the main network (the fifth in our setup). Its objective is to identify their field of origin – source or target. In a min-max training paradigm, the main network is trained, besides estimating redshift, to produce feature maps that will fool the adversarial network.

The adversarial module is trained with the cross-entropy loss function (Equation (1)), which quantifies the disparity between its predictions,  $p_i$ , and the domain labels,  $d_i$ , defined as 0 for the source and 1 for the target:

$$L = - \sum_{i=1}^2 d_i \log(p_i). \quad (1)$$

This loss function updates only the adversarial layers. Following the min-max paradigm, the main layers of the network, which supply the feature map input to the adversarial module, were trained with the inverse of this loss, known as the confusion loss. It was computed by just inverting the domain labels,  $d_i$ . Thus, these layers aim to generate feature maps that maximize the adversarial error, while enabling accurate redshift predictions. The adversarial and redshift estimation objectives were optimized simultaneously at each iteration, using the same learning rate and the same cross-entropy loss function.

As this network carries a risk of negative domain transfer (by producing indistinguishable feature maps for objects that should be distinguishable, such as a faint source domain galaxy and a bright target domain galaxy), we added two complementary features:



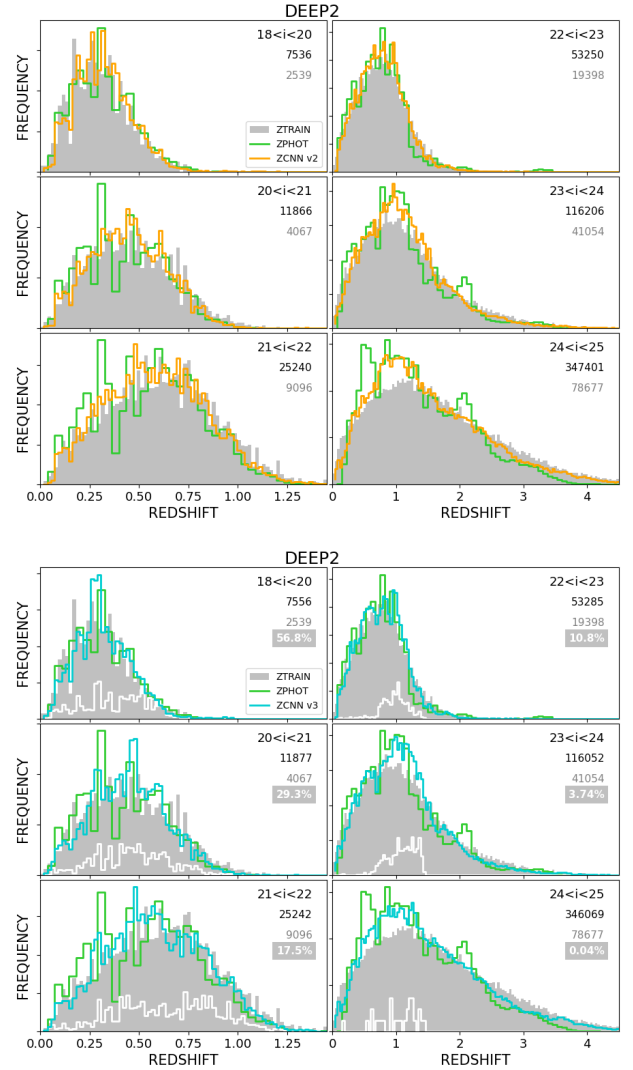
**Fig. 11.** Performance of the adversarial model trained with the COSMOS UDF sample alone (v2, orange lines) or with the COSMOS UDF sample supplemented by the DEEP2 spectroscopic sample in cross-validation mode (v3, turquoise lines) for the DEEP2 spectroscopic sample as a function of redshift and magnitude. The dark blue and pink lines are the baseline results (v0 and v1) shown in Fig. 8. The *ugrizy* SED-fitting metrics are shown in green.

- 1 the selection of the target batch based on magnitude: negative domain transfer can be avoided by using pseudo labels guiding the adversarial training to align similar objects in the two domains (Pei et al. 2018). We chose the *i*-band magnitude, which may help disentangle the imaging conditions from the physical properties that are crucial for redshift estimation. In practice, for each galaxy in the source batch, we randomly selected one galaxy from the target domain among the 30 sources that are closest in *i*-band magnitude.
- 2 pairing the selections: to make the magnitude pairing idea more efficient, we modified the DA learning objective. Instead of having the adversarial network determine whether feature maps originate from the source or target domain, it had to decide if they came from the same field or not. To do so, we concatenated the feature maps along the width axis before feeding them to the adversarial network. The adversarial network produced one estimate per pair of images. We ensured that 50% of the pairs are from the same field (equally split between target and source) and the remaining 50% are pairs from different domains.

## 5.2. Results

We trained the adversarial CNN with the COSMOS UDF training sample as source data and 100 000 unlabeled DEEP2 sources as target data. The PDFs inferred on DEEP2 are the average of ten trainings. We refer to this model as v2. We also trained the adversarial model using the COSMOS UDF sample supplemented by the spectroscopic DEEP2 data as a training set with a cross-validation protocol. We refer to this last experiment as v3.

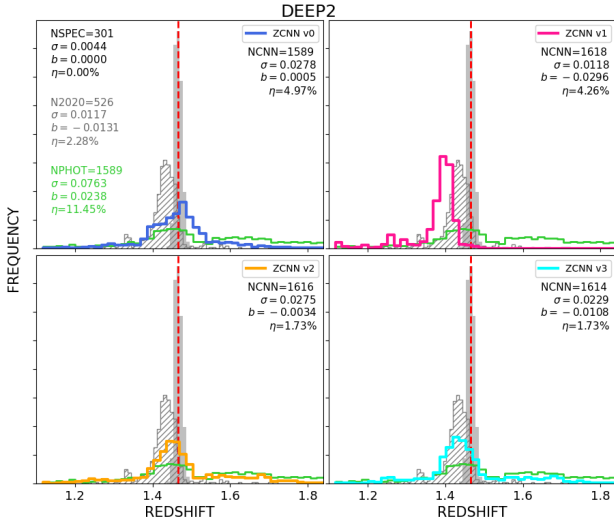
Figure 11 shows the v2 and v3 metrics as a function of redshift and magnitude for the DEEP2 spectroscopic sample, compared to the baseline results (v0 and v1) shown in Fig. 8 and to the SED-fitting metrics. The v2 versus v0 comparison, the two models trained on COSMOS UDF alone, demonstrates the significant impact of the adversarial module. Adding the DEEP2



**Fig. 12.** Predicted redshift distributions of the full DEEP2 data inferred from the adversarial network trained on the COSMOS UDF sample alone (v2, orange histograms in the top panels) and on the COSMOS UDF data supplemented by the DEEP2 spectroscopic data (v3, turquoise histograms in the bottom panels). See Fig. 9 for the full caption.

spectroscopic data to the training set brings the v3 metrics nearly down to the v1 level.

Contrary to the baseline results in Fig. 9, a dramatic improvement is also seen in the redshift distributions inferred for the unlabeled galaxies, shown in Fig. 12. These distributions are significantly more consistent with the reference sample at faint magnitude, without or with the addition of the DEEP2 spectroscopic data to the training sample. The two models are very similar but the v3 model appears to create more structures, most notably at  $i < 20$  and at  $i > 23$ . This suggests that the presence of even a small fraction of unrepresentative target sources amid the training data may still induce ill-understood biases despite the adversarial module, and that the improved metrics for the spectroscopic target population remain misleading. The most conspicuous discrepancy with the underlying training distribution for both models is an excess of galaxies around  $z \sim 1$ , taken away from the low- and high-redshift tails. This may be a flaw, but it is also conceivable that we over-flattened the COSMOS UDF distribution (Fig. 5) and that this growth is at least partly real. This point is further discussed in Section 7.

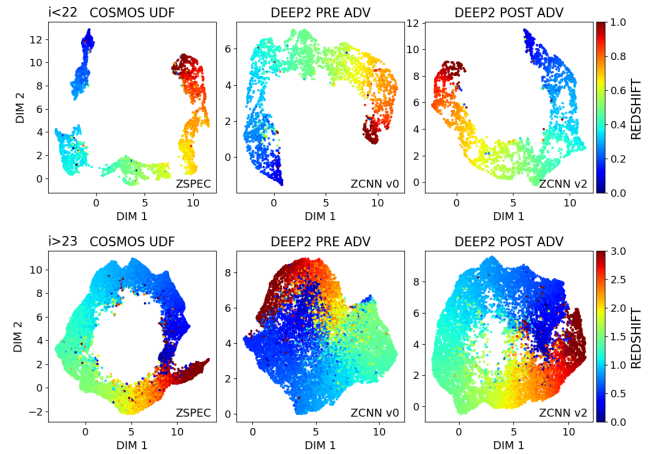


**Fig. 13.** Redshifts inferred from the v0, v1, v2, and v3 models for a sample of [OII] emission line galaxy candidates at  $z \approx 1.47$  in the DEEP2 field. Their *ugrizy* SED-fitting redshift distribution is shown in green. The 526 C2020 redshifts available for such galaxies in the COSMOS field are shown as hatched histograms and the spectroscopically confirmed cases in all the fields combined as gray-filled histograms. The corresponding metrics, in the top right corner of each panel for the CNN, and in green and gray for the *ugrizy* and C2020 redshifts respectively, assume a true redshift  $z_{EL} = 1.4657$  (red vertical line, see text for details).

A final test is afforded by a sample of  $\sim 15\,000$  galaxies selected from narrow-band observations designed to detect [OII] emission line galaxies at redshift  $z \approx 1.47$  (Section 2.2). This sample spans a large magnitude range and covers COSMOS, DEEP23, ELAIS, and bits of XMM. Figure 13 shows the redshifts predicted for the  $\sim 1600$  such galaxies at  $i < 25$  in the DEEP2 region by the v0, v1, v2, and v3 models. The distributions are compared to the *ugrizy* SED-fitting redshift distribution and to that of the 526 C2020 redshifts available for the emission line candidates in the COSMOS UDF. The corresponding metrics quoted in the panels assume a true redshift  $z_{EL}$  of 1.4657. We computed  $z_{EL}$  as the median redshift of the 301 most reliable spectroscopic redshifts available for these galaxies at  $i < 25$  in all the fields combined, assuming  $ZFLAG > 2$  and redshifts within the  $x$  axis range of the plots,  $1.1 < z < 1.84$ , which corresponds to  $|\Delta z| < 0.15$ , our definition of acceptable estimates (non-catastrophic failures). We note that five spectroscopic redshifts with  $ZFLAG > 2$  are in this category, signaling that the narrow-band filter detections are not 100% [OII] emission line galaxies at  $z \approx 1.47$ . All values of  $\eta$  are, by definition, the fraction of redshift estimates that extend beyond the plots.

Interestingly, the baseline model, v0, although ruled out, does not do such a bad job. The baseline model trained with the addition of the DEEP2 spectroscopic sources, v1, also ruled out, induces a strong negative bias, confirming its inadequacy. Both versions of the adversarial model, v2 and v3, yield significantly more decent results, in the sense that they align with the C2020 histogram, fulfilling the goal of the adversarial module since C2020 redshifts dominate the training sample. The fraction of catastrophic failures is also much reduced. Like v1, v3, whose training also includes the DEEP2 spectroscopic sources, induces a stronger negative bias.

Given this test and the shapes of the overall redshift distributions, we settle for v2 as our preferred model. UMAP



**Fig. 14.** UMAP embeddings from a deep convolutional layer, colored by redshift. The leftmost panels displays a COSMOS UDF validation sample from the baseline training, colored by redshift labels. The middle and right panels show the embeddings from the same layer for DEEP2 unlabeled galaxies, colored by redshift predictions, before and after DA (v0 vs. v2), respectively. The top and bottom rows make use of galaxy samples with  $i < 22$  and  $i > 23$ , respectively.

(McInnes et al. 2018) or t-SNE (van der Maaten & Hinton 2008) can help us visualize the impact of DA within our network. To select a relevant layer, we used a COSMOS UDF validation sample from the trained baseline network (Section 4.2) to quantify the quality of the internal representations produced at each layer. For every layer, we extracted per-galaxy activation vectors, applied dimensionality reduction using UMAP and t-SNE, and computed silhouette scores with respect to redshift. As expected, the deepest layers exhibit the strongest feature separability in redshift space. The top and bottom left panels of Fig. 14 show the UMAP embeddings from one of the last convolutional layers (more visually dispersed than its t-SNE equivalent) for validation galaxies with  $i < 22$  and  $i > 23$ , respectively, colored by redshift (spectroscopic or C2020 labels). The four right-hand panels show UMAP embeddings from the same layer for a random subset of bright and faint unlabeled DEEP2 galaxies colored by predicted redshift, before and after DA (v0 vs. v2). The improvement in redshift encoding at faint magnitude is significant, confirming the results presented in this section. We note that the lowest and highest redshifts are not perfectly separated in the COSMOS UDF validation sample at  $i > 23$ , a weakness that DA cannot improve upon. The fact that labels are predominantly C2020 photometric estimates at faint magnitude, with low- and high-redshift outliers (visible in Fig. 3 for the bright-magnitude biased sample with spectroscopic redshifts), contributes to degrading redshift classification as magnitude increases (this is further discussed in Section 7).

## 6. The other HSC regions

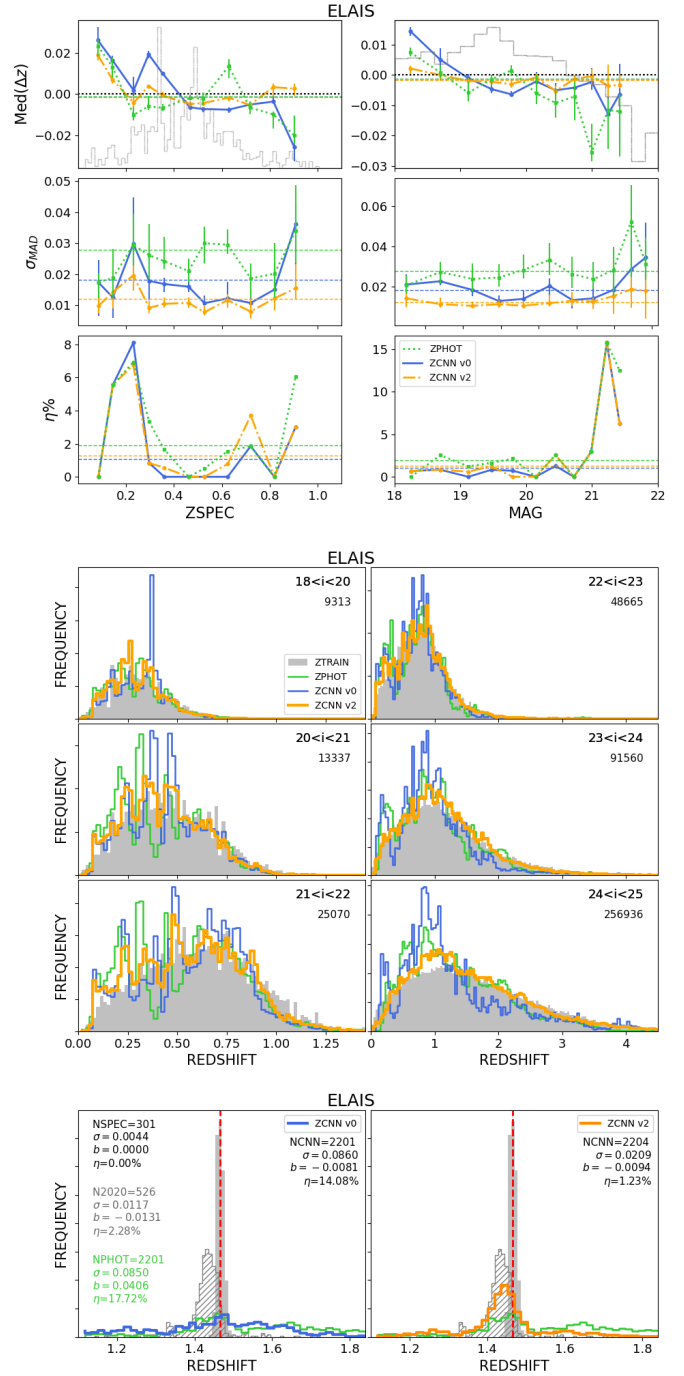
In the following subsections, we present the results of our adaptation model applied to the four other HSC subregions<sup>4</sup>: ELAIS, XMM ULTRA DEEP, XMM DEEP, and COSMOS DEEP. As described in Section 2.1, each of these fields has its own photometric characteristics, which more or less differ from the

<sup>4</sup> For each region we performed ten training runs of 50 epochs using V100 and A100 GPUs on the Jean Zay supercomputer at IDRIS. One training run consumes  $\sim 17$  hours of V100 or  $\sim 7.5$  hours of A100, with a memory footprint of  $\sim 27$  GB (RES).

COSMOS UDF, making adaptation more or less critical for redshift estimation based on that particular source field. The quality of the SED-fitting photometric redshifts also differs: in XMM DEEP, XMM ULTRA DEEP, and COSMOS DEEP, they were computed with the addition of three NIR bands for 51.8%, 24.6%, and 3% of the sources, respectively. In the case of COSMOS DEEP, 71.6% of the “spectroscopic” sample are C2020 redshifts.

The results of the adaptation model (v2) for these four subregions are compared to the inferences from the baseline CNN (v0) and to the SED-fitting redshifts, in Figs. 15, 16, 17 and 18. These figures show, from top to bottom, the performance of the three methods for the spectroscopic samples, their predicted redshift distributions for the complete samples in six magnitude bins, and their predicted redshift distributions for the [OII] emission line galaxy candidates.

- ELAIS (Fig. 15): the ELAIS field is very poor in spectroscopy, with only  $\sim 1200$  redshifts lower than 1 in this field. The adversarial model is not a dramatic improvement on the baseline CNN or even on the *ugrizy* SED-fitting redshifts for those few essentially bright, low-redshift sources. The redshift distributions inferred from the v0 and v2 models for the full sample at bright magnitude are also similarly distorted by unlikely prominent redshift structures. Only at  $i > 23$  does the v2 model appear to be a significant improvement on v0 and on the SED-fitting redshift. An excess of faint galaxies around  $z \sim 1$  with respect to the training distribution is present as in DEEP2 but less prominently. The emission line galaxy predictions are brought considerably closer to the C2020 values.
- XMM ULTRA DEEP (Fig. 16): the improvement from SED-fitting to v0 to v2 is readily visible in this field. The v2 redshift distributions are still quite distorted at bright magnitude but there are three times fewer galaxies than in ELAIS and DEEP2. The SED-fitting photometric redshifts were computed with the addition of three NIR bands (Section 2.1) for a small quarter of the sources, which, strangely, seems to degrade the redshift distributions at faint magnitude and to cause the emission line galaxy predictions to be more negatively biased than in the DEEP2 and ELAIS cases. The v0 model is more aligned with spectroscopically confirmed emission line galaxies than v2 but the latter has learned what it was designed for.
- XMM DEEP (Fig. 17): the difference between v0 and v2 is less significant in this field for the spectroscopic sources, but the gain from v2 is once again obvious in the redshift distributions of the complete sample. The distributions at  $i > 23$  show the largest excess of galaxies around  $z \sim 1$  of all the HSC regions. The emission line galaxy predictions are also the worst of all fields but the sample is much smaller. XMM DEEP has the lowest S/N of all regions (Fig. 1), which probably contributes to the degraded results (P19). We also found that its  $r - i$  color distribution is shifted by  $\sim -0.2$  compared to the other fields, while all other colors align. Some flaw in the calibrations may also contribute.
- COSMOS DEEP (Fig. 18): this region surrounds the COSMOS UDF. Its labeled sample contains 8561 spectroscopic redshifts and 6359 C2020 redshifts on the high-S/N outskirts of the UDF. The S/N remains quite high throughout most of the region (Fig. 1). In this particular case, the difference between v0 and v2 is understandably minimal compared to the other regions, including for the redshift distributions of the complete sample. Yet the adaptation module still



**Fig. 15.** Predictions of the baseline CNN (v0 in blue) and adversarial CNN (v2 in orange) trained with the COSMOS UDF sample for the ELAIS data, compared to the *ugrizy* SED-fitting redshifts (in green). *Top*: metrics measured for the spectroscopic sample as a function of redshift and magnitude. *Middle*: predicted redshift distributions of the complete catalog in six magnitude bins. See Fig. 9 for the complete caption. *Bottom*: redshifts inferred for the [OII] emission line candidates. See Fig. 13 for the complete caption.

provides a noticeable improvement at the faintest magnitudes and for [OII] emission line candidates.

The most notable discrepancy between the predicted redshift distributions of the complete samples in all regions and the training distributions is an excess of galaxies around  $z \sim 1$  in

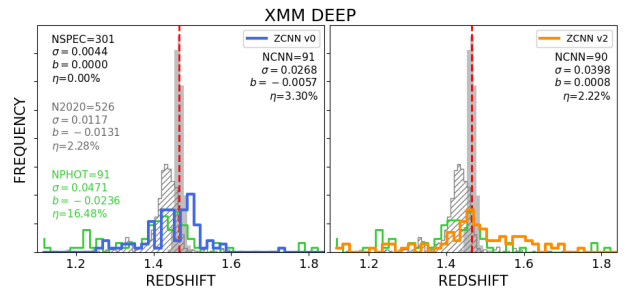
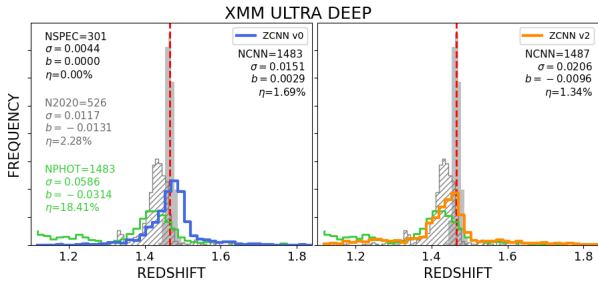
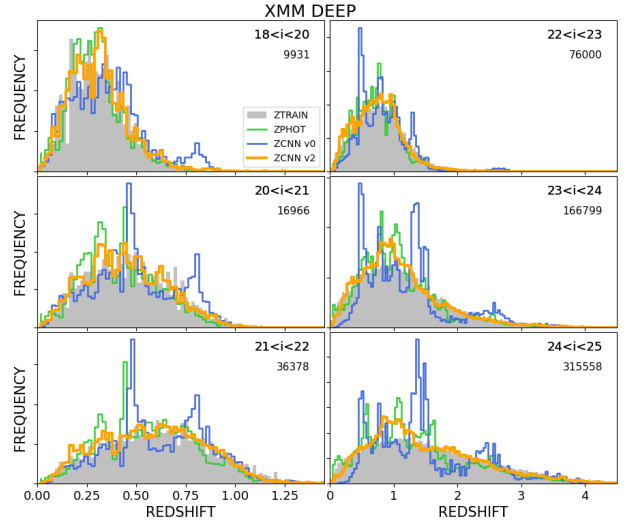
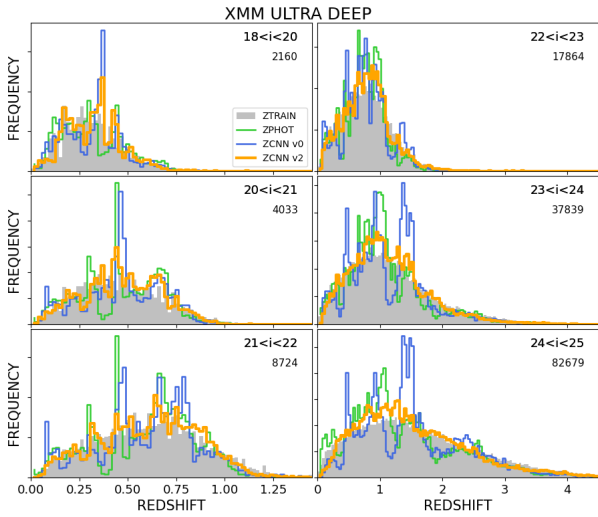
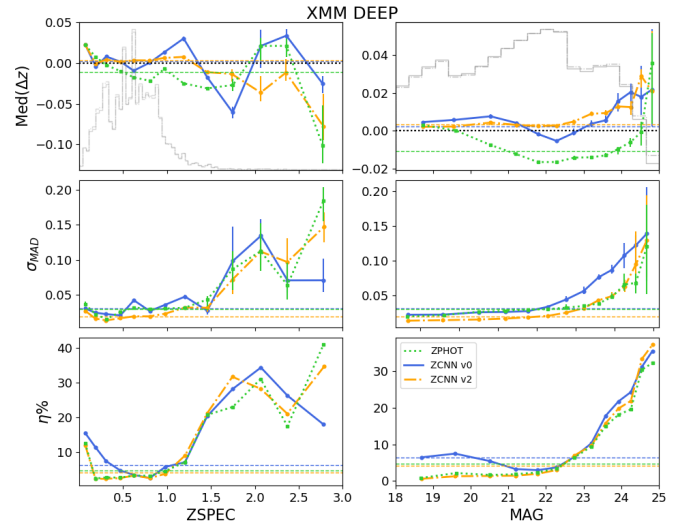
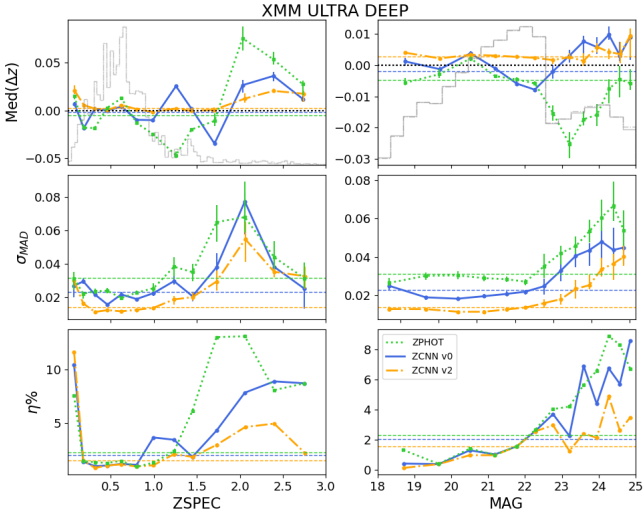


Fig. 16. Same as Fig. 15, but for the XMM UDF.

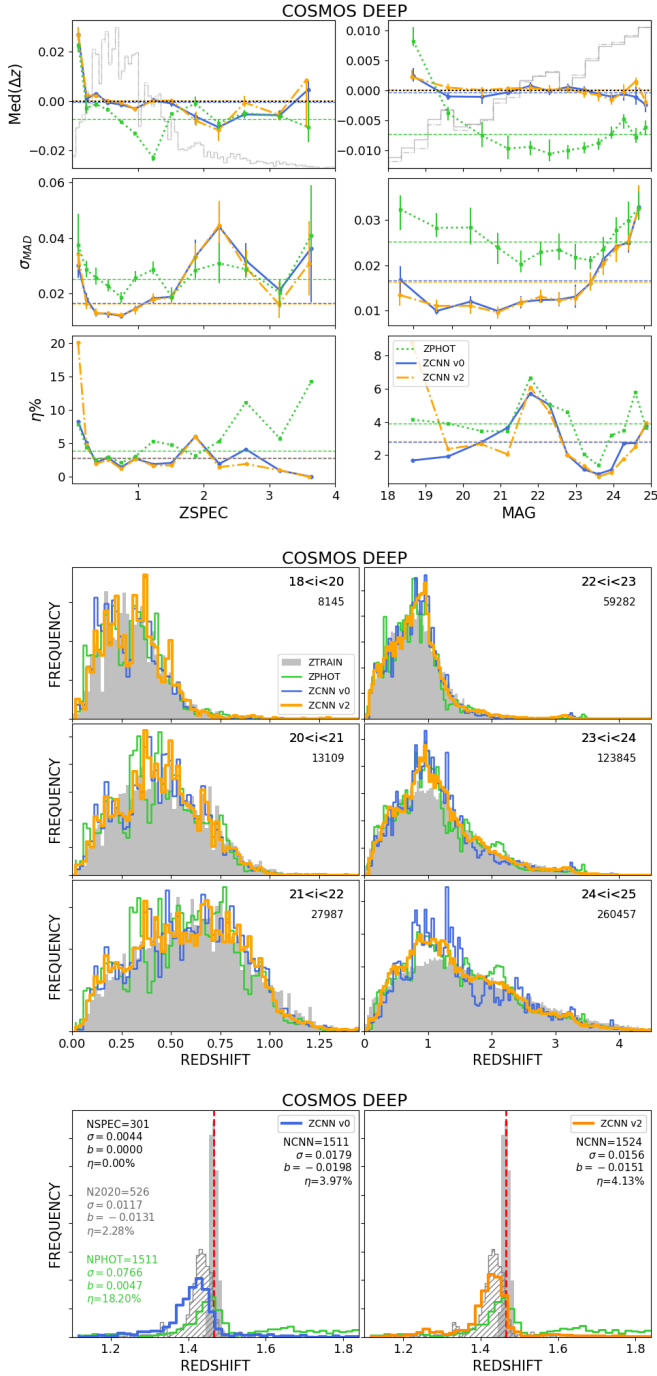
Fig. 17. Same as Fig. 15, but for the XMM DEEP field.

magnitude, at the expense of the low- and high-redshift tails. This excess is a flawed result of the model insofar as we consider the training distributions to be representative, but, as noted in Section 5.2, our smoothing of the COSMOS UDF redshift distribution (Fig. 5) may not be the best representation of the sky-averaged redshift distribution of galaxies. Figure 19 overlays the predicted redshift distributions of the five regions in the faintest magnitude bin,  $24 < i < 25$  (shown in previous figures), with the original COSMOS UDF distribution and its smoothed training version. Even if other regions of the sky have no reason to exhibit such big double structure around  $z \sim 1$ , our smoothing procedure could very well underestimate the actual frequency of galaxies in that redshift range, in which case the network's predictions may not be so far from the ground truth. It is actually a

success not to reproduce the training distribution, which should ideally be completely flat.

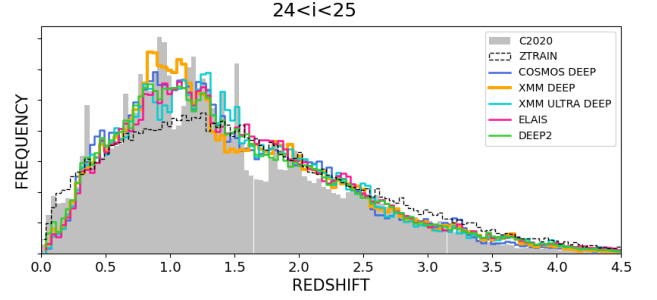
## 7. Quality assessment

A good assessment of redshift reliability is provided by the width of the CNN classification, PDFw, defined as the redshift interval underlying the central 68% of the PDF (Pasquet et al. 2019; Treyer et al. 2024). Figure 20 shows the v2 PDFw distribution of the complete samples at  $i < 25$  in the 5 HSC regions analyzed in the previous sections. All fields show a resurgence of very insecure predictions at PDFw > 1.5, emphasized by

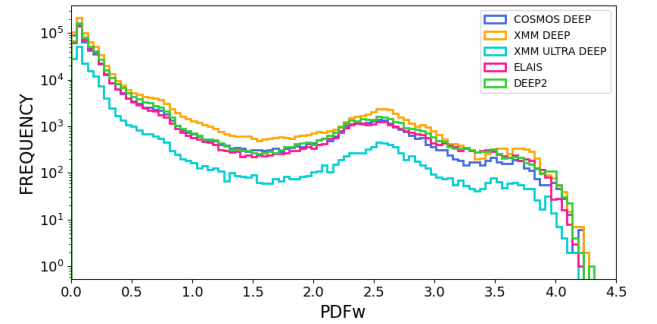


**Fig. 18.** Same as Fig. 15, but for the COSMOS DEEP field.

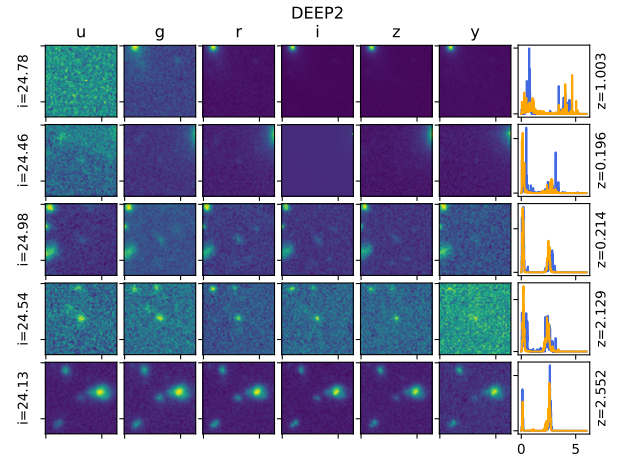
the log scale on the  $y$  axis. These cases increase with magnitude, reaching between 7.3% and 9.5% at  $24 < i < 25$ . The PDFw distributions of the v0 model have a similar shape, but with fractions of PDFw  $> 1.5$  ranging from 10.7% to 22.7% at  $24 < i < 25$ , bounded by XMM UDF and XMM DEEP. Figure 21 shows a random sample of DEEP2 galaxies with v2 PDFw  $> 1.5$ , to be compared to samples with v2 PDFw  $< 1.5$  shown in Appendix A. The images appear of poor quality and/or have multiple sources brighter than the central target, but that is not unusual in Fig. A.1 either. Both the v0 and v2 PDFs are wide due to their multimodality. In the examples displayed in Appendix (v2 PDFw  $< 1.5$ ), v2 seems to have resolved the v0 multimodal conflicts. In all that follows, we restrict the statistics to PDFw  $< 1.5$ .



**Fig. 19.** Predicted (v2) redshift distributions in the 5 HSC regions at  $24 < i < 25$ , compared to the original COSMOS UDF distribution (shaded gray histogram) and its smoothed training version (dashed line). All histograms are normalized to 1.

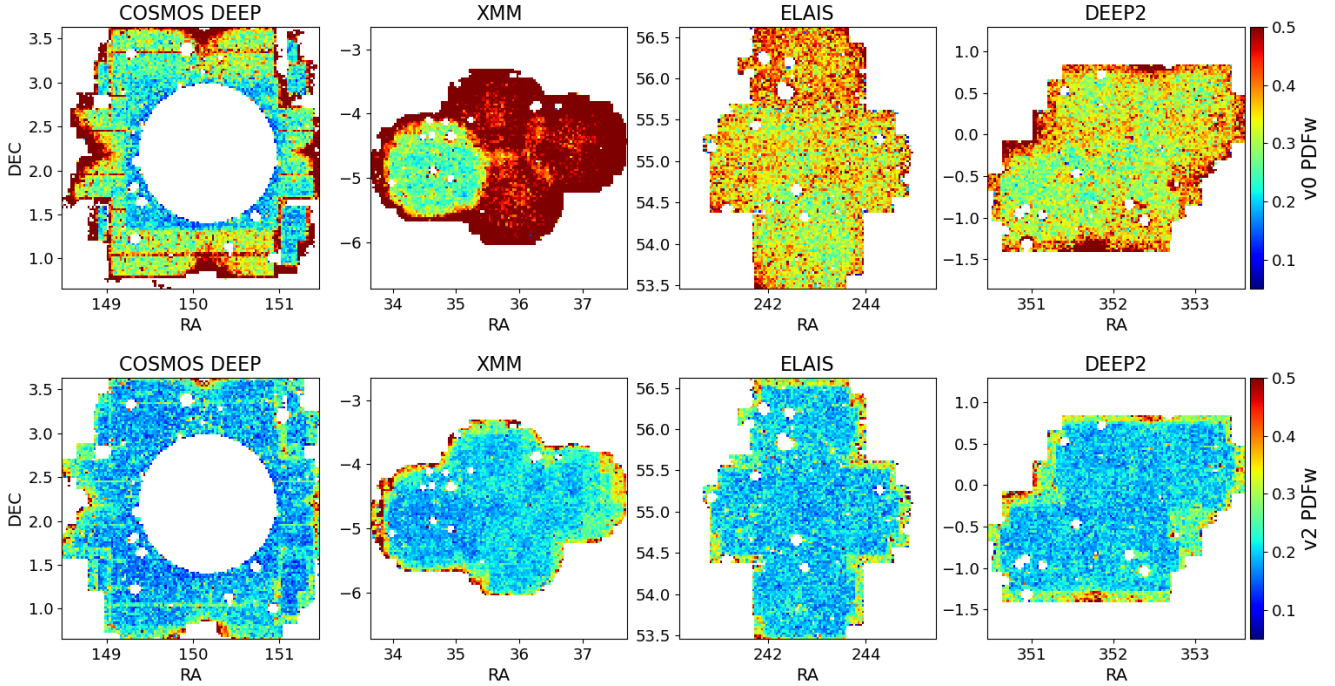


**Fig. 20.** PDF width distribution of the complete samples in the five HSC regions (v2 model).

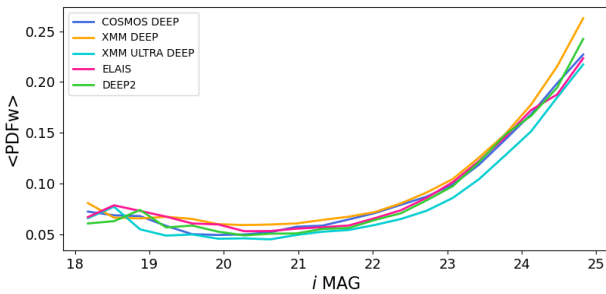


**Fig. 21.** Random sample of  $ugrizy$  images of DEEP2 galaxies at  $24 < i < 25$  with v2 PDFw  $> 1.5$ , to be compared to samples with PDFw  $< 1.5$  in Fig. A.1. The right most panels show the PDFs from the v0 and v2 models in blue and orange, respectively. The redshift values are the final v2 estimates.

Figure 22 shows the gain in PDFw per region from the baseline CNN to the adversarial model at  $24 < i < 25$ . Note that since PDFw larger than 1.5 have been excluded in both cases, more poor values have been discarded in the upper panels than in the lower ones (more than twice as many in the case of XMM DEEP). The gain is therefore larger than it looks. XMM DEEP clearly has its own specific issues compared to the other regions. As mentioned in Section 6, we found a color shift in  $r - i$  color compared to the other fields, and to COSMOS UDF in particular, which might explain why the baseline inference fails more dramatically than in other regions. The adversarial model largely



**Fig. 22.** The four HSC regions color-coded by the PDF width (PDFw < 1.5) of sources at  $24 < i < 25$  for the v0 and v2 models in the upper and lower panels, respectively. The color scale is the same for all.



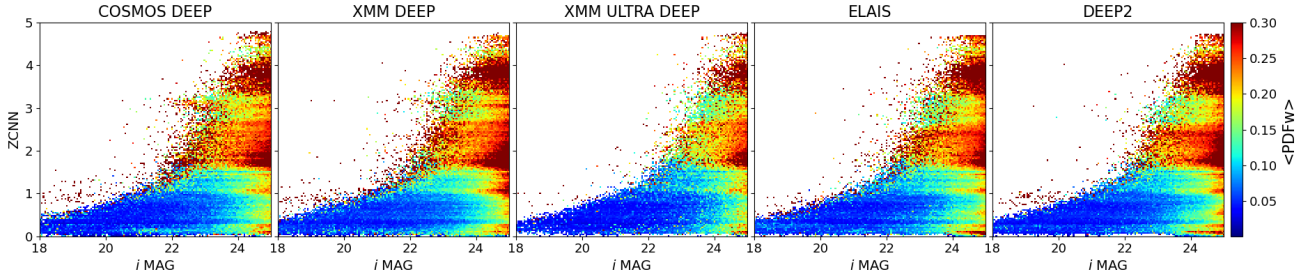
**Fig. 23.** Mean PDF width (PDFw < 1.5) as a function of magnitude in the five HSC regions (v2 model). The range is bounded by XMM ULTRA DEEP and XMM DEEP, which have the lowest and largest values, respectively.

remedies this problem. The XMM UDF, which is closest to the COSMOS UDF in S/N and E(B-V), fares consistently better. Poor results persist on the outskirts of the fields.

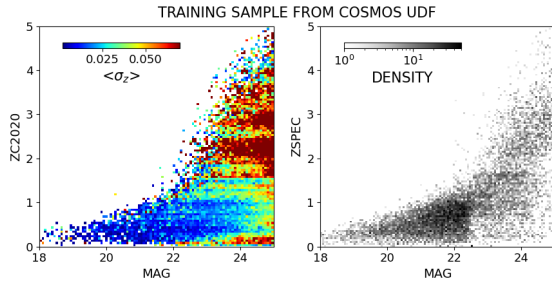
Although the mean PDF width increases smoothly with magnitude (Fig. 23), the degradation in redshift accuracy exhibits some complexity. Figure 24 shows the mean PDF width per pixel in magnitude-redshift space in the five regions. The patterns are remarkably similar in all the fields. At  $i \gtrsim 23$ , the widening of the PDFs (essentially due to multimodality) varies unevenly with redshift (their median values), with two extended patches of particularly degraded predictions around  $z_{\text{CNN}} \approx 1.8$  and  $z_{\text{CNN}} \approx 3.8$ . In contrast, the network appears much less hesitant about piling up galaxies around  $z \sim 1$ , which aligns with the earlier argument that the predicted redshift frequency in this range may not be a classification failure. XMM DEEP, whose “excess” around  $z \sim 1$  is more extreme than in the other regions (Fig. 19), is more degraded at this particular redshift (Fig. 24).

As mentioned in Section 2.2, the C2020 redshifts we use as labels are the means of four estimates, with dispersion  $\sigma_z \leq 0.1(1 + \bar{z})$ . The left panel of Fig. 25 shows the mean  $\sigma_z$  per pixel in the magnitude-redshift space. It appears that  $\sigma_z$  also varies unevenly with redshift at faint magnitudes. Although the red patches are not centered around the same redshifts as in Fig. 24, the topological similarity between the distributions in both figures is noteworthy. It suggests that the CNN detects the more or less blurry nature of the redshift labels depending on their location in the parameter space and that it affects its classification ability, reflected by the PDF widths. The right panel shows the magnitude-redshift distribution of the spectroscopic redshifts we substituted to the C2020 values in the training sample, when available. At  $i > 23$ , they account for less than 8% of the sample. They may be too few to have much stabilizing influence on the classification or, as they tend to be larger than the C2020 estimates (the small negative bias in Fig. 3 worsens with redshift), they may add to the CNN’s confusion around these galaxies. In any case, we may assume that the uneven consistency of the labels reflected by the  $\sigma_z$  distribution is at least partly responsible for the uneven PDFw distribution in Fig. 24, even if not straightforwardly.

All distributions in Fig. 24 plateau at  $z \sim 4.8$ , above which no point estimate is ever predicted despite training the model to  $z = 6$ . The latter phenomenon was also observed when deriving photometric redshifts in the SDSS via a CNN classification (P19). It appears to be due to the shortage of galaxies in the high-redshift tail of the training distribution, which does not adequately populate the last bins of the classification. Wider bins may be used at the highest redshifts to alleviate the problem (Treyer et al. 2024). At faint magnitudes, the network is also plagued by other weaknesses of the training sample, in particular the lower S/N, and of the inference samples, whose S/N are even lower, except in the XMM UDF.



**Fig. 24.** Distributions of PDF uncertainties (mean PDF width per pixel) in magnitude-redshift space. They are remarkably similar in all fields, with similarly uneven patches of particularly degraded predictions at  $i \gtrsim 23$ .



**Fig. 25.** *Left:*  $\sigma_z$  distribution of the C2020 labels (see Section 2.2) in magnitude-redshift space. *Right:* spectroscopic redshifts replacing C2020 values in the training sample. In the red regions on the left, the CNN relies essentially on dubious labels.

## 8. Conclusion

We used a representative sample of  $\sim 180\,000$  redshifts available in the COSMOS UDF as labels to train a CNN with the aim of predicting redshifts for the remaining  $\sim 4$  million galaxies in the HSC-CLAUDS survey at  $i < 25$  (Desprez et al. 2023). For most of the brightest galaxies in this sample ( $i < 22.5$ ), the redshifts are spectroscopic; for most of the faintest galaxies, they are photometric, based on 30-band SED-fitting from the COSMOS2020 sample (Weaver et al. 2022). The input data supplied to the CNN consist of six stamp images in the *ugrizy* bands centered on each galaxy of the HSC survey. The CNN architecture is from Ait Ouahmed et al. (2024). Its output is a classification into contiguous bins that we use as a redshift PDF, whose median and width are our point estimate and uncertainty measure, respectively.

Training the CNN with the COSMOS-UDF labeled sample revealed a very strong domain mismatch problem when inferring redshifts for sources from other HSC regions. To alleviate this problem, we developed an unsupervised adversarial-based DA network. This approach aims to transfer the redshift classification capability that the CNN learns from the COSMOS UDF data – the source field – to another so-called “target” field. Its effectiveness was tested with: 1, the metrics that can be measured for the spectroscopic sample available in each target field (mostly bright galaxies); 2, the shape of the magnitude-dependent redshift distributions predicted for the entire target population, which are expected to resemble those of the presumably representative training sample; and 3, the redshift predictions for a sample of galaxies detected by deep, narrow-band filter observations expected to catch strong [OII] emitters at  $z \sim 1.47$ . All three tests demonstrate that the adversarial network is able to transfer its ability to estimate redshifts from the source training set to the other regions. It has yielded a unified representation that captures the underlying data structure. We find the PDF widths

to be good measures of redshift reliability. In particular, they are able to point back to the regions of the magnitude-redshift space where the labels are most uncertain.

The limited quality of the training sample remains a significant downside. The training redshifts from COSMOS2020, or COSMOS2025 (Shuntov et al. 2025) for a future analysis at fainter magnitudes, may be better represented by soft instead of one-hot labels. The method could also be combined with self-supervised techniques further ensuring that the resulting latent space does not discriminate between the different fields. New self-supervised methods for extracting robust features from large databases have emerged in astrophysics, often relying on foundation models (Parker et al. 2024; Lastufka et al. 2024). However, these models remain biased, and t-SNE projections can end up separating data by field or survey rather than revealing true intrinsic structure. Despite their flaws, our results show that adversarial-based DA networks are capable of remedying this bias by aligning representations across different regions. Our approach offers promising perspectives for redshift estimation in future large imaging surveys in which little spectroscopy will be available for training, such as those expected from the Vera Rubin telescope. Pending improvements, the redshift catalogs we generated for the COSMOS, XMM, DEEP2, and ELAIS fields are available at [deepdip.iap.fr](http://deepdip.iap.fr).

*Acknowledgements.* This work was carried out using computing and storage resources at IDRIS thanks to grants 2024-AD010414147R1 and 2025-AD010414147R2 awarded by GENCI on the V100 and A100 partitions of the Jean Zay supercomputer. It benefited from the support of the French National Research Agency (ANR) as part of the DEEPDIP project (ANR-19-CE31-0023).

## References

- Aihara, H., Arimoto, N., Armstrong, R., et al. 2018, *PASJ*, 70, S4
- Aihara, H., AlSayyad, Y., Ando, M., et al. 2019, *PASJ*, 106
- Ait Ouahmed, R., Arnouts, S., Pasquet, J., Treyer, M., & Bertin, E. 2024, *A&A*, 683, A26
- Alexander, S., Gleyzer, S., Parul, H., et al. 2023, *ApJ*, 954, 28
- Belfiore, F., Ginolfi, M., Blanc, G., et al. 2025, *A&A*, 694, A212
- Bertin, E., Mellier, Y., Radovich, M., et al. 2002, in *Astronomical Society of the Pacific Conference Series*, 281, Astronomical Data Analysis Software and Systems XI, eds. D. A. Bohlender, D. Durand, & T. H. Handley, 228
- Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., & Krishnan, D. 2017, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3722
- Carliles, S., Budavári, T., Heinis, S., Priebe, C., & Szalay, A. S. 2010, *ApJ*, 712, 511
- Ćiprijanović, A., Kafkes, D., Downey, K., et al. 2021, *MNRAS*, 506, 677
- Ćiprijanović, A., Kafkes, D., Snyder, G., et al. 2022, *Mach. Learn. Sci. Technol.*, 3, 035007
- Ćiprijanović, A., Lewis, A., Pedro, K., et al. 2023, *Mach. Learn. Sci. Technol.*, 4, 025013
- Collister, A. A., & Lahav, O. 2004, *PASP*, 116, 345
- Csabai, I., Budavári, T., Connolly, A. J., et al. 2003, *AJ*, 125, 580

- Desprez, G., Picouet, V., Moutard, T., et al. 2023, *A&A*, **670**, A82
- Dey, B., Andrews, B. H., Newman, J. A., et al. 2022, *MNRAS*, **515**, 5285
- D’Isanto, A., & Polsterer, K. L. 2018, *A&A*, **609**, A111
- Farrens, S., Lacan, A., Guinot, A., & Vitorelli, A. Z. 2022, *A&A*, **657**, A98
- Ganin, Y., Ustinova, E., Ajakan, H., et al. 2016, *J. Mach. Learn. Res.*, **17**, 2096
- Ghifary, M., Kleijn, W. B., Zhang, M., & Balduzzi, D. 2015, in *Proceedings of the IEEE International Conference on Computer Vision*, 2551
- Gilda, S., de Mathelin, A., Bellstedt, S., & Richard, G. 2024, *Astronomy*, **3**, 189
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al. 2014, *Adv. Neural Inform. Process. Syst.*, **27**
- Hayashi, M., Shimakawa, R., Tanaka, M., et al. 2020, *PASJ*, **72**, 86
- Hayat, M. A., Stein, G., Harrington, P., Lukić, Z., & Mustafa, M. 2021, *ApJ*, **911**, L33
- Henghes, B., Pettitt, C., Thiyagalingam, J., Hey, T., & Lahav, O. 2021, *MNRAS*, **505**, 4847
- Henghes, B., Thiyagalingam, J., Pettitt, C., Hey, T., & Lahav, O. 2022, *MNRAS*, **512**, 1696
- Hoyle, B. 2016, *Astron. Comput.*, **16**, 34
- Huertas-Company, M., Iyer, K. G., Angeloudi, E., et al. 2024, *A&A*, **685**, A48
- Ilbert, O., Arnouts, S., McCracken, H. J., et al. 2006, *A&A*, **457**, 841
- Jarvis, M. J., Bonfield, D. G., Bruce, V. A., et al. 2013, *MNRAS*, **428**, 1281
- Khostovan, A. A., Kartaltepe, J. S., Salvato, M., et al. 2025, arXiv e-prints [arXiv:2503.00120]
- Lastufka, E., Bait, O., Taran, O., et al. 2024, *A&A*, **690**, A310
- Li, C., Zhang, Y., Cui, C., et al. 2022, *MNRAS*, **509**, 2289
- Li, C., Zhang, Y., Cui, C., et al. 2024, *AJ*, **168**, 233
- Long, M., Cao, Y., Wang, J., & Jordan, M. 2015, in *International Conference on Machine Learning*, PMLR, 97
- McCracken, H. J., Milvang-Jensen, B., Dunlop, J., et al. 2012, *A&A*, **544**, A156
- McInnes, L., Healy, J., Saul, N., & Großberger, L. 2018, *J. Open Source Softw.*, **3**, 861
- Menou, K. 2019, *MNRAS*, **489**, 48026
- Parker, L., Lanusse, F., Golkar, S., et al. 2024, *MNRAS*, **531**, 4990
- Pasquet, J., Bertin, E., Treyer, M., Arnouts, S., & Fouchez, D. 2019, *A&A*, **621**, A26
- Pei, Z., Cao, Z., Long, M., & Wang, J. 2018, *Multi-Adversarial Domain Adaptation*
- Picouet, V., Arnouts, S., Le Floc’h, E., et al. 2023, *A&A*, **675**, A164
- Roncoli, A., Čiprijanović, A., Voetberg, M., Villaescusa-Navarro, F., & Nord, B. 2023, arXiv e-prints [arXiv:2311.01588]
- Roster, W., Salvato, M., Krippendorf, S., et al. 2024, *A&A*, **692**, A260
- Sawicki, M., Arnouts, S., Huang, J., et al. 2019, *MNRAS*, **489**, 5202
- Shuntov, M., Akins, H. B., Paquereau, L., et al. 2025, arXiv e-prints [arXiv:2506.03243]
- Swierc, P., Tamargo-Arizmendi, M., Čiprijanović, A., & Nord, B. D. 2024, arXiv e-prints [arXiv:2410.16347]
- Treyer, M., Ait Ouahmed, R., Pasquet, J., et al. 2024, *MNRAS*, **527**, 651
- Tzeng, E., Hoffman, J., Saenko, K., & Darrell, T. 2017, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7167
- van der Maaten, L., & Hinton, G. 2008, *J. Mach. Learn. Res.*, **9**, 2579
- Vanzella, E., Cristiani, S., Fontana, A., et al. 2004, *A&A*, **423**, 761
- Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2020, *Nat. Methods*, **17**, 261
- Wadadekar, Y. 2005, *PASP*, **117**, 79
- Weaver, J. R., Kauffmann, O. B., Ilbert, O., et al. 2022, *ApJS*, **258**, 11
- Wei, S., Li, C., Zhang, Y., et al. 2025, *PASA*, **42**, e092
- Xu, Q., Shen, S., de Souza, R. S., et al. 2023, *MNRAS*, **526**, 6391
- Yao, L., Qiu, B., Luo, A. L., et al. 2023, *MNRAS*, **523**, 5799
- Ye, R., Shen, S., de Souza, R. S., et al. 2025, *MNRAS*, **537**, 640
- York, D. G., Adelman, J., Anderson, Jr., J. E., et al. 2000, *AJ*, **120**, 1579
- Zhang, Y., Ma, H., Peng, N., Zhao, Y., & Wu, X.-b. 2013, *AJ*, **146**, 22
- Zhang, C., Wang, W., Qu, M., Jiang, B., & Zhang, Y. 2024, *AJ*, **168**, 244

## Appendix A: Images and PDFs

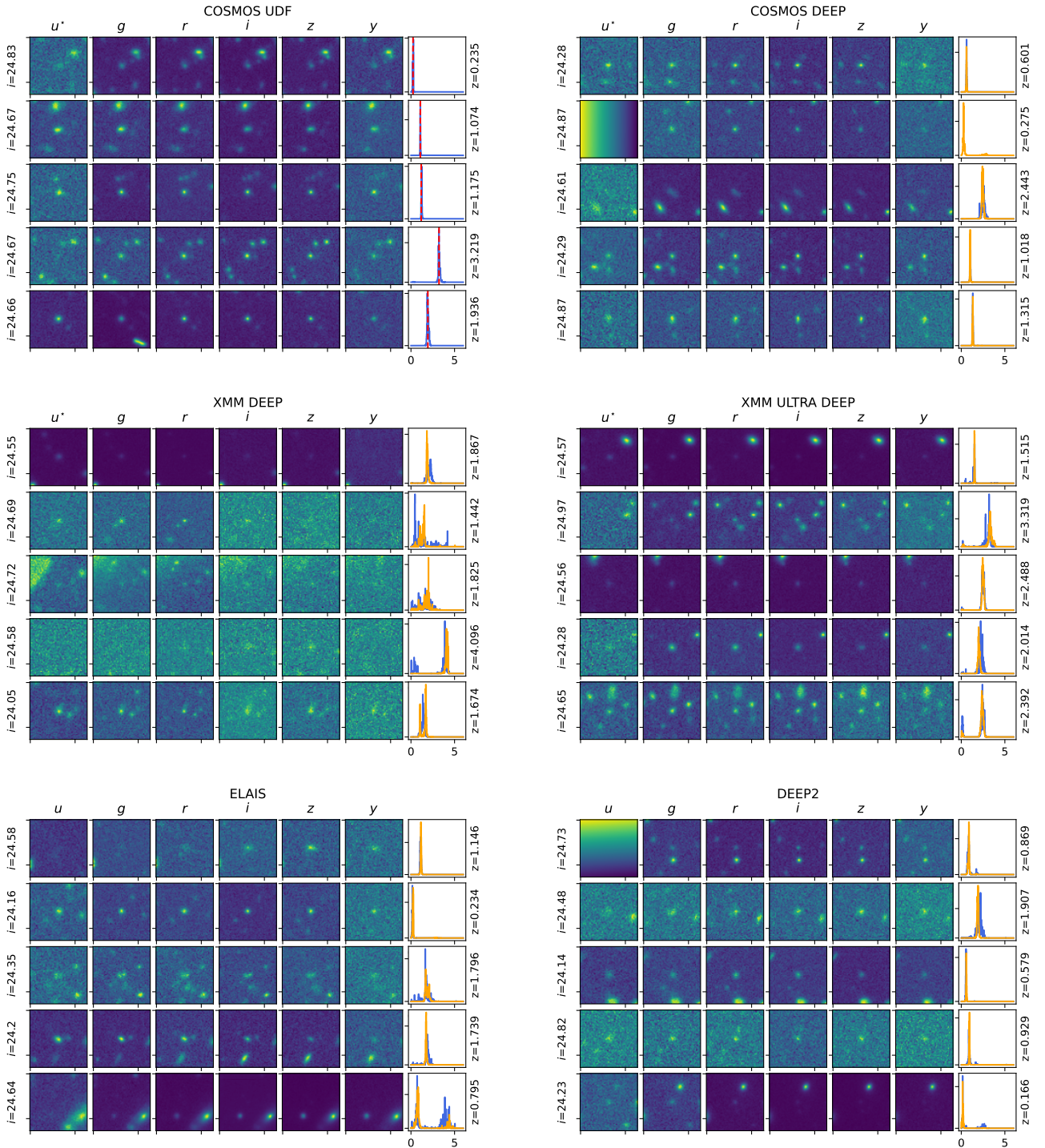


Fig. A.1: A random sample of  $u^*grizy$  or  $ugrizy$  images of galaxies at  $24 < i < 25$  in the 6 HSC subregions. For the COSMOS UDF, the right most panels show the PDFs from the baseline cross-validation training and the  $z$  values on the right, marked by vertical red dashed lines, are the redshift labels (spectroscopic or C2020). For the regions other than COSMOS UDF, the right most panels show the PDFs (PDFw < 1.5) from the v0 model in blue, and for the v2 model in orange. The  $z$  values on the right are the v2 redshift estimates.