Deep learning in quasar physics

F. Rastegar Nia

Physics Department, Alzahra University Vanak, 1993891176, Tehran, Iran
ICRA and Dipartimento di Fisica, Università di Roma "La Sapienza", Piazzale Aldo Moro 5, I-00185 Roma, Italy
ICRANet, Piazza della Repubblica 10, I-65122 Pescara, Italy
E-mail: f.rastegarnia@alzahra.ac.ir

M. T. Mirtorabi

Physics Department, Alzahra University Vanak, 1993891176, Tehran, Iran
E-mail: torabi@alzahra.ac.ir

R. Moradi* and Y. Wang†

ICRANet, Piazza della Repubblica 10, Pescara,I–65122, Italy,
ICRA, Dipartimento di Fisica, Sapienza Universita' di Roma, Rome, I–00185, Italy and INAF, Rome, 00136, Italy
**E-mail: ruffini@icra.it*
**E-mail: rahim.moradi@icranet.org*
†*E-mail: yu.wang@icranet.org*

A. Vafaei Sadr

Departement de Physique Theorique and Center for Astroparticle Physics, University of Geneva
E-mail: Alireza.VafaeiSadr@unige.ch

In view of increasing data volume of existing and upcoming telescopes/detectors we here apply the 1–dimensional convolutional neural network (CNN) to estimate the redshift of (high-)redshifts quasars in Sloan Digital Sky Survey IV (SDSS-IV) quasar catalog from DR16 of eBOSS. Our CNN takes the flux of the quasars as an array and their redshift as labels. We here evidence that new structure of the network, and augmenting the training set, provide a high precision result in estimating the redshift of quasars.

Keywords: Quasar, Deep learning, CNN, SDSS

1. Introduction

Quasars are the most luminous active galactic nuclei(AGN) which are powered by accretion disk around supermassive black holes at centers of their host galaxies. Thanks to their high luminosity, they can be observed across the universe in wide range of redshift from z = 0 to z ∼ 7. Thus, quasars give us important information about the early universe, the structure formation and evolution.[1,2]

Nowadays, astronomy and astrophysics have been brought into the big data era through the construction and development of ground-based and space telescope. Since huge amount of data, ranging from gamma-ray, x-ray, ultraviolet, optical,

infrared and radio bands of quasars is available in sky surveys, various challenges and opportunities are created for scientific discoveries.[3,6]

In recent years, machine learning (ML) and deep learning (DL) have been utilized in astronomy and astrophysics in order to deal with the big data surveys and as well extract the new physical understanding. The goal in ML and DL is to diagnose, by optimization, common characteristics and features in data.[10] These interesting branch of computer science have been used in estimating the physical parameters and classification of celestial objects; for example the morphological classification of galaxies,[15,16,20] estimation of photometric redshift,[12,25,33,34] classification of star/galaxy,[4,27,32] spectral classification of stars.[5,7,17,35]

This article is mainly dedicated to train a CNN to predict the redshift of Quasi-stellar radio source (quasars) or quasi stellar objects (QSO) in Sloan Digital Sky Survey IV (SDSS-IV) quasar catalog from Data Release 16 (DR16) of the extended Baryon Oscillation Spectroscopic Survey (eBOSS), which is the most comprehensive catalog of spectroscopically collected quasars to date.[26]

2. Dataset

In this paper, we exploit dataset from the quasar spectra obtained by the Sloan Digital Sky Survey-4(SDSS-4). They are provided by the Sixteenth Data Release (DR16) of SDSS extended Baryon Oscillation Spectroscopic Survey (eBOSS).[14] According to the technical details of eBOSS, namely the wavelength coverage in range of 361nm$-$1014nm with the resolution $R = 2000$, where $R = \frac{\lambda}{\Delta\lambda}$, more than 700,000 quasar spectra have been detected in the redshift range $0 \leq z \leq 7.1$.

The distributions of detected quasars redshift in the DR16 catalog is shown in Fig. 1. As can be seen, the number of quasars reaches its maximum around $z \approx 2.5$; at earlier epochs i.e., higher redshifts, they are comparatively rare.

2.1. *Redshift*

The observed quasar spectra contain broad emission lines, often time-variable flux both in the continuum and in the emission lines and UV flux; see e.g.[11] The shift and deformation of the various lines and specific characteristic of different fluxes in the spectra of quasars, due to the cosmological redshift, makes quasars a proper tool to invest the deep learning method and extract the common features hidden in their spectra (flux).

In DR16Q catalog, redshifts obtained for quasars by different methods, such as principal component analysis (PCA) and QuasarNET are reported. In this work we use "*primary*" redshift or best redshift for quasars which has been selected from the available visual inspection redshifts,[8] or, alternatively, the SDSS automated pipeline redshift;, see[23,24] and references therein for more information. The best option is exploiting only the visual inspection redshifts which will be presented in the forthcoming papers.

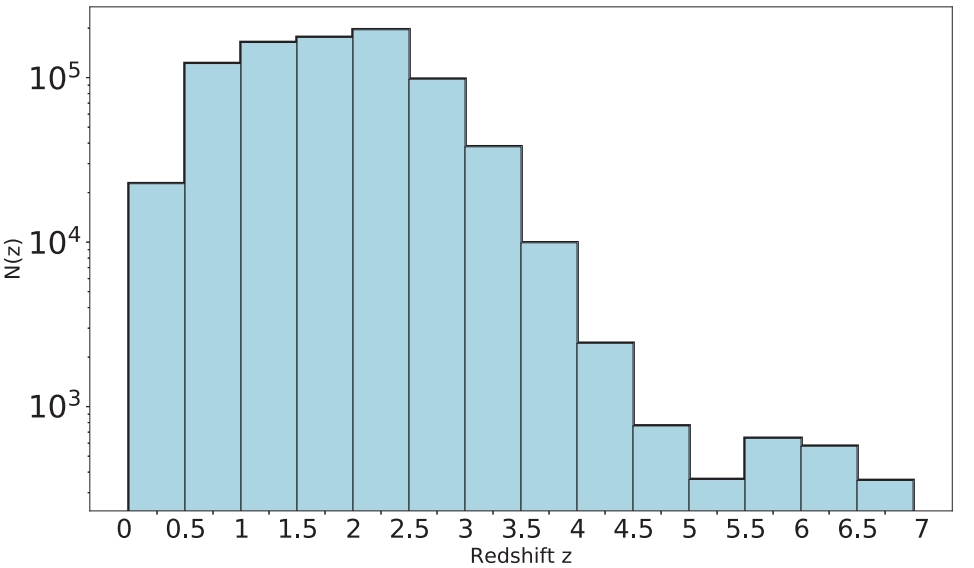

Fig. 1. The distributions of "*primary*" redshift of quasars in the DR16 catalog.

3. Convolutional neural network (CNN)

3.1. *CNN architecture*

Since the spectra of quasars are represented by flux versus the wavelength, namely 1+1 dimensional set of data, they can be considered as as time-series. For this reason, the neural network model in this work is designed to be a 1-dimensional CNN. It operates like a combination of some mathematical functions, which optimizes through several filters, and transforms the input data; here flux of quasars, to the output data, here redshift of quasars, by extracting the hidden features from the quasars' fluxes.

The physical information, such as redshift are hidden in the observed flux, and CNNs, as a regressor, are supposed to estimate the redshift through the training process. The CNN in the current work is a combination of convolutional and fully-connected layers. The convolutional layers are initialized with He Normal initializer.[22] If the model requires to access the non-linear modes in data he Rectified Linear Unit (ReLU) activation function is implemented.[36]

The filters in each convolutional layer scans each row (flux) and extract the prominent features from the raw input for the specific redshift. The feature-extraction layers have a repeating pattern of the convolutional layers whereas the pooling layers reduce the dimensions and concentrate only on the most prominent features. This process recognizes the non-linear correlation within the dataset. Finally, the fully-connected feed-forward layers connect and assign the extracted

features to the output layer (redshift). All free parameters in the model change dynamically as the algorithm finds the best solution, achieved by the back-propagation learning algorithm.

Many free parameters, such as the number of layers, layer specifications, and their arrangement can be changed while one constructs a neural network model; known as model architecture. The architecture selection affects optimization and quality of the performance; this process is called hyperparameter optimization.[18] The hyper-parameter optimization process is necessary in order to ensure that neither underfits nor overfits of the training set happens.

In the hyper-parameter optimization process, the number and length of filters of convolutional layers, the number of nodes in fully-connected layers, and the kernel size of the Maxpooling layer are considered and finally the "Mean Squared Error" (MSE), as the loss function, is utilized in our CNN.

We also use Adam optimizer to optimize the loss function, which is an algorithm for the first-order gradient-based optimization of stochastic objective functions. We here set the learning rate as 0.0001 and the coefficients for computing the averages of gradient and its square as 0.5 and 0.9, the weight decay of L2 penalty is set as 0.[28, 29]

Figure 2 demonstrates the CNN pipeline of this work. It takes a quasar spectrum as a 1-dimensional array and predicts the redshift. We have tested different samples of training set, and the addition of more layers to this configuration does not enhance the prediction accuracy presented in this paper for different redshift intervals.

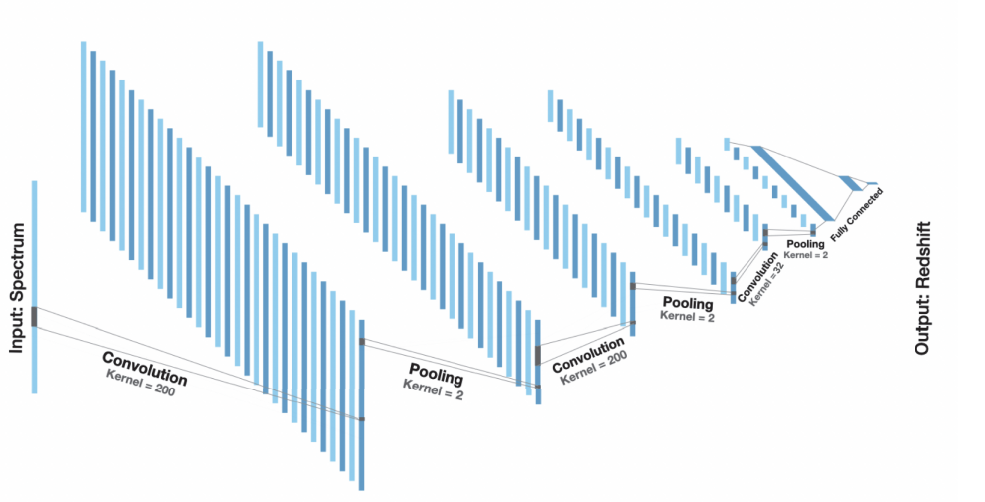

Fig. 2. Structure of 1-dimensional CNN developed in this work to learn higher-order features hidden in the input spectrum. The CNN goes through the spectrum via a convolutional layer of kernel size = 200, 200, 32, respectively in order to search for the global and local pattern. The fully connected layers output the redshift.

3.2. *Pre-processing the data*

Data pre–processing is crucial in two aspects: (1) to provide an understandable dataset for DL networks, and (2) to increase the speed and accuracy of processing. As the first step of the pr–eprocessing procedure in the present work, a two-dimensional matrix of the dataset is created. In the matrix each row represents the quasar flux and each column relates to a flux a certain wavelength. At the second step, the flux of every spectrum is normalized via the Zero-Mean Normalization method. Moreover, the normalized fluxes are stored in a Numpy array and applied as features dataset, their corresponding redshifts are collected for labels dataset as well.

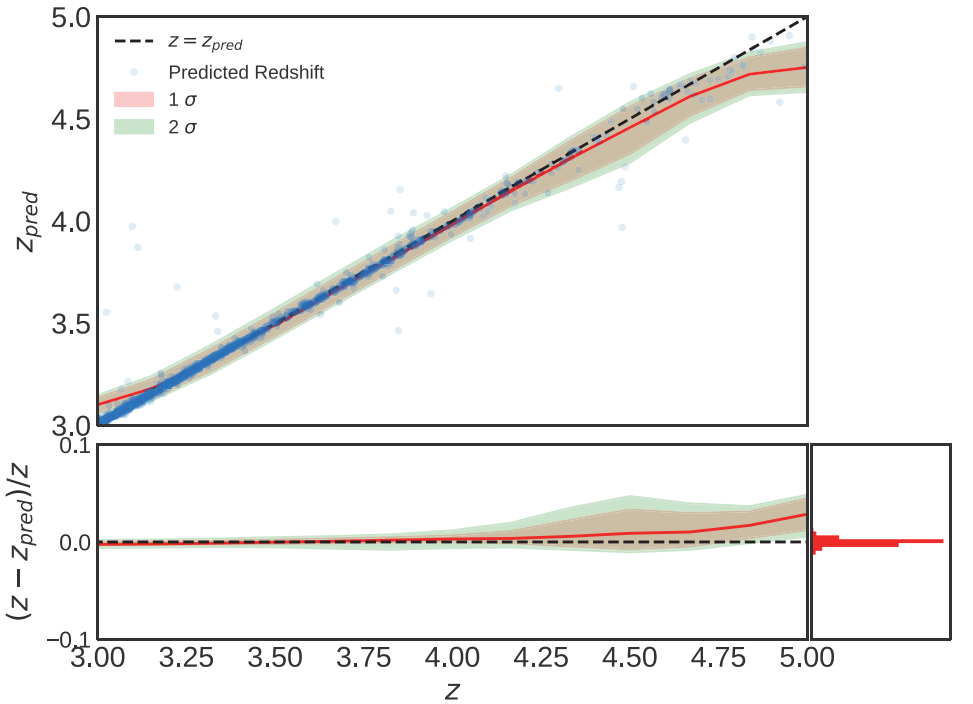

Fig. 3. Redshift predicted by our CNN for $3 \leq z \leq 5$. Top: The predicted redshift vs. the redshift reported in the DR catalog. Bottom: The relative error and their distribution. As it can be seen the distribution of relative error follows a Gaussian distribution with mean, $\mu = 0.0042$ and a standard deviation of $\sigma = 0.014$.

4. Prediction in the $3 \leq z \leq 5$ interval

As an example, we use the quasars in the $3 \leq z \leq 5$ interval as sources to test our CNN. In general, there are 51, 100 quasars in this interval,75% of them, namely 38832 spectra are taken as the sample in order to train our CNN (training set) and

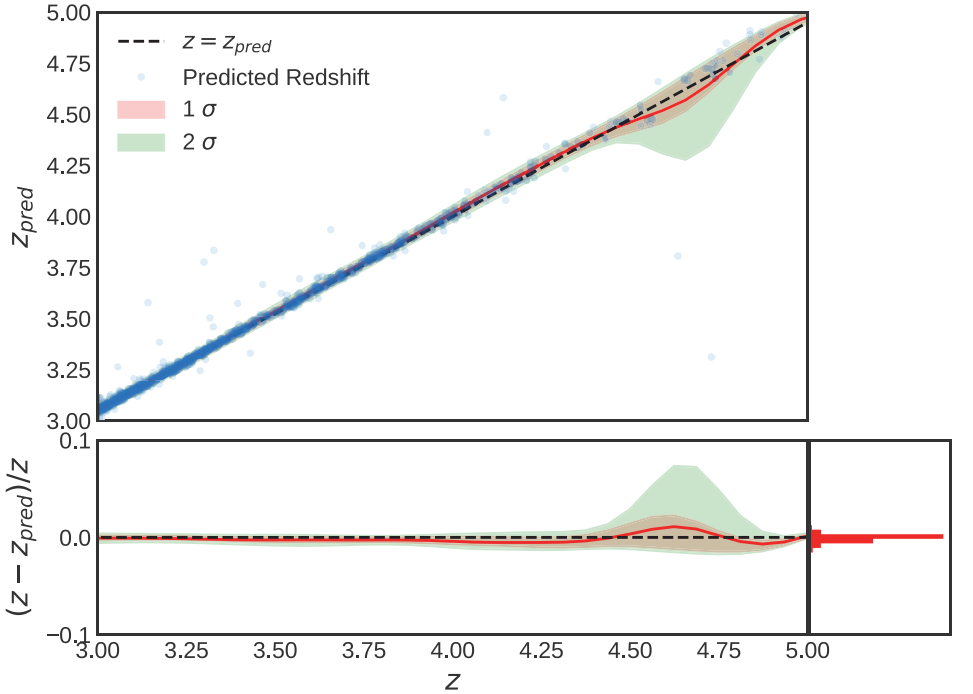

Fig. 4. Redshift predicted by our CNN for $3 \leq z \leq 5$ with training set in $2.9 \leq z \leq 5.2$ interval. Top: The predicted redshift vs. the redshift reported in the DR catalog. Bottom: The relative error and their distribution. The distribution of relative error follows a Gaussian distribution with mean, $\mu = 0.0038$ and a standard deviation of $\sigma = 0.012$.

25%, namely 12776 as sources to testify our CNN (test set). In the whole paper, the test set has not been used to train the CNN. Moreover, no limit on the S/N ratio of spectra is imposed.

The predicted redshift by our CNN for $3 \leq z \leq 5$ together with their relative errors distribution and the best fit is shown in Fig. 3. The distribution of relative error, RE,

$$\mathrm{RE} \equiv \frac{\mathrm{z} - \mathrm{z}_{\mathrm{pred}}}{\mathrm{z}}, \tag{1}$$

which z is the redshift reported by SDSS and z_{pred} is the redshift predicted by our CNN, follows a Gaussian distribution with mean, $\mu = 0.0042$ and a standard deviation of $\sigma = 0.014$. The method used for smoothing the scatterplot is locally weighted regression (LWR).[13]

As it can be seen in Fig. 3, the fit in the lower and upper edges of the prediction, namely $4.5 \lesssim z \leq 5$ and $3 \lesssim z \leq 3.1$ intervals, is not as significant as other time intervals.

This problem can be due to:

(1) the lack of data in $z < 3$ and $z > 5$ and consequently the failed LWR in the aforementioned intervals.
(2) the deficiency of training sample in $4.5 \lesssim z \leq 5$.

In order to solve the first issue, we extend our training set to $2.9 \leq z \leq 5.2$. By adding this extra data to the train set we make sure that there is no bias in the LWR fitting presented in Fig. 3 at its lower and upper part; see Fig. 4.

In order to accommodate the second issue and increase the number of spectra of the training set in $4.5 \lesssim z \leq 5$ we first convert the observed spectra in $4 \leq z \leq 4.5$ interval, where there are enough spectra to produce a reliable training set, into their cosmological rest-frame by dividing their wavelength by 1+z then we convert data set to $4.5 \leq z \leq 5$ by multiplying the wavelengths by a factor of 1+z+0.5.

Figure 5 represents the results when extra samples from $2.9 \leq z \leq 3$ and $5 \leq z \leq 5.2$ as well as the redshifted sample from $4 \leq z \leq 4.5$ to $4.5 \leq z \leq 5$ have been added to the original training set. The distribution of relative error follows a

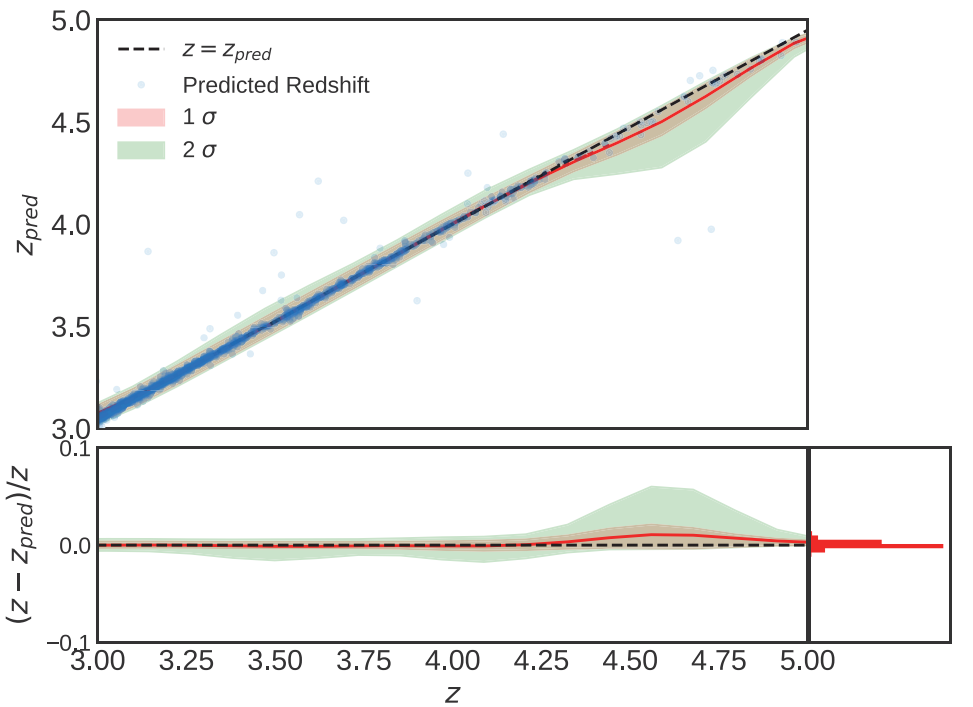

Fig. 5. Redshift predicted by our CNN for $3 \leq z \leq 5$ with training set in $2.9 \leq z \leq 5.2$ interval together with the converted observed spectra in $4 \leq z \leq 4.5$ interval into $4.5 \leq z \leq 5$. Top: The predicted redshift vs. the redshift reported in the DR catalog. Bottom: The relative error and its distribution which follows a Gaussian distribution with mean, $\mu = 0.0030$ and a standard deviation of $\sigma = 0.010$.

Gaussian distribution with mean, $\mu = 0.003$ and a standard deviation of $\sigma = 0.01$. The prediction is clearly improved in $4.5 \leq z \leq 5$ comparing to initial result shown in Fig. 3.

5. Conclusions

Deep learning, especially CNN[31,37] is promising in the future Astrophysical studies. CNN can be utilized, owing to its several convolutional and fully connected layers to find the common deep hidden patterns in the spectrum, in provision for improving our astrophysical knowledge of distant objects. In this regard, we showed that exploiting a CNN leads to a statistically significant prediction of the redshift of quasars. We also showed that augmentation of dataset, which depends on the statistical and physical features of samples improves the prediction of redshift. Since CNN finds the physical characteristics of the spectra and these characteristics are deformed by the cosmological redshift the results presented here can be extended to the other energy bands like, X-ray, Infrared, UV, Gamma-ray and etc.

References

1. Nina Hernitschek et al 2016 ApJ 817 73, DOI: https://doi.org/10.3847/0004-637X/817/1/73.
2. A. Lupi, M. Volonteri, R. Decarli, S. Bovino, J. Silk and J. Bergeron, Mon. Not. Roy. Astron. Soc. **488** (2019) no.3, 4004-4022 doi:10.1093/mnras/stz1959 [arXiv:1901.02464 [astro-ph.GA]].
3. Allen, G., Andreoni, I., Bachelet, E., et al. 2019, arXiv preprint arXiv:1902.00522.
4. Bai, Y., Liu, J., Wang, S., & Yang, F. 2018, The Astronomical Journal, 157, 9.
5. Bailer-Jones, C. A., Irwin, M., & Von Hippel, T. 1998, Monthly Notices of the Royal Astronomical Society, 298, 361.
6. Ball, N. M., & Brunner, R. J. 2010, International Journal of Modern Physics D, 19, 1049.
7. Bialek, S., Fabbro, S., Venn, K. A., et al. 2019, arXiv preprint arXiv:1911.02602.
8. Bolton, A. S., Schlegel, D. J., Aubourg, É., et al. 2012, AJ, 144, 144, doi:10.1088/0004-6256/144/5/144.
9. Busca, N., & Balland, C. 2018, arXiv e-prints, arXiv:1808.09955.
10. Carleo, G., Cirac, I., Cranmer, K., et al. 2019, Rev. Mod. Phys., 91, 045002, doi:10.1103/RevModPhys.91.045002.
11. Carroll, B. W., & Ostlie, D. A. 1996, An Introduction to Modern Astrophysics.
12. Cavuoti, S., Brescia, M., Tortora, C., et al. 2015, Monthly Notices of the Royal Astronomical Society, 452, 3100.
13. Cleveland, W. S. 1979, Journal of the American Statistical Association, 74, 829, doi:10.1080/01621459.1979.10481038.
14. Dawson, K. S., Kneib, J.-P., Percival, W. J., et al. 2016, The Astronomical Journal, 151, 44.
15. De La Calleja, J., & Fuentes, O. 2004, Monthly Notices of the Royal Astronomical Society, 349, 87.
16. Dobrycheva, D., Vavilova, I., Melnyk, O., & Elyiv, A. 2017, arXiv preprint arXiv:1712.08955.

17. Fabbro, S., Venn, K., O'Briain, T., et al. 2018, Monthly Notices of the Royal Astronomical Society, 475, 2978.
18. Feurer, M., & Hutter, F. 2019, in Automated Machine Learning (Springer, Cham), 3–33.
19. Fiorentin, P. R., Bailer-Jones, C., Lee, Y. S., et al. 2007, Astronomy & Astrophysics, 467, 1373.
20. Gauci, A., Adami, K. Z., & Abela, J. 2010, arXiv preprint arXiv:1005.0390.
21. Glazebrook, K., Offer, A. R., & Deeley, K. 1998, The Astrophysical Journal, 492, 98.
22. He, K., Zhang, X., Ren, S., & Sun, J. 2015, in Proceedings of the IEEE international conference on computer vision, 1026–1034.
23. Hewett, P. C., & Wild, V. 2010, MNRAS, 405, 2302, doi:10.1111/j.1365-2966.2010.16648.x.
24. Higley, A. N., Lyke, B. W., Myers, A. D., et al. 2020, in American Astronomical Society Meeting Abstracts, Vol. 235, American Astronomical Society Meeting Abstracts #235, 219.03.
25. Hoyle, B. 2016, Astronomy and Computing, 16, 34.
26. Hutchinson, T. A., Bolton, A. S., Dawson, K. S., et al. 2016, AJ, 152, 205, doi:10.3847/0004-6256/152/6/205.
27. Kim, E. J., & Brunner, R. J. 2016, Monthly Notices of the Royal Astronomical Society, stw2672.
28. Kingma, D. P., & Ba, J. 2014, arXiv preprint arXiv:1412.6980.
29. LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. 1998, Proceedings of the IEEE, 86, 2278.
30. Li, X.-R., Pan, R.-Y., & Duan, F.-Q. 2017, Research in Astronomy and Astrophysics, 17, 036.
31. Liu, C.-L., Hsaio, W.-H., & Tu, Y.-C. 2018, IEEE Transactions on Industrial Electronics, 66, 4788.
32. Odewahn, S. C., Stockwell, E., Pennington, R., Humphreys, R. M., & Zumach, W. 1992, in Digitised Optical Sky Surveys (Springer), 215–224.
33. Pasquet-Itam, J., & Pasquet, J. 2018, A& A, 611, A97, doi:10.1051/0004-6361/201731106.
34. Sadeh, I., Abdalla, F. B., & Lahav, O. 2016, Publications of the Astronomical Society of the Pacific, 128, 104502.
35. Sharma, K., Kembhavi, A., Kembhavi, A., et al. 2020, Monthly Notices of the Royal Astronomical Society, 491, 2280.
36. Xu, B., Wang, N., Chen, T., & Li, M. 2015, arXiv preprint arXiv:1505.00853.
37. Xu, L., Ren, J. S., Liu, C., & Jia, J. 2014, Advances in neural information processing systems, 27, 1790.