A Global to Local Guiding Network for Missing Data Imputation

WeiWang[1,3,4] Yimeng Chai[2,3,4] Yue Li[1,3,4,*]

[1] Key Laboratory for Medical Data Analysis and Statistical Research of Tianjin (KLMDASR)

[2] Tianjin Key Laboratory of Network and Data Security Technology

[3] Trusted AI Laboratory, College of Cyber Science, Nankai University, Tianjin, China

[4] College of Computer Science, Nankai University

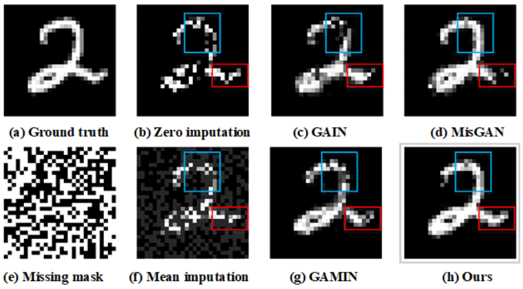
ABSTRACT

Missing data imputation aims to accurately impute the unobserved regions with complete data in real world. Although many recent methods have made remarkable advances, the local homogenous regions especially in boundary and the reasonable of the imputed data are still two most challenging issues. To address these issues, we propose a novel Global to Local Guiding Network (G2LGN) based on generative adversarial network for missing data imputation, which is composed of a Global-Impute-Net (GIN), a Local-Impute-Net (LIN) and an Impute Guider Model (IGM). The GIN looks at the entire missing regions to generate and impute data as a whole. Considering the reasonable of the GIN results, IGM is assigned to capture coherent information between global and local and guide the LIN to look only at a small area centered at the missing focused regions. After the processing of these three modules, local imputed results are concatenated to global imputed results, which impute reasonable values and refine local details from rough to accurate. The comprehensive experiments on both numeric datasets and image dataset demonstrate our method is significantly superior to other 3 state-of-the-art approaches and 7 traditional methods. Besides, the extensive ablation study validates the superior performance for dealing with missing data imputation.

Index Terms— Missing data imputation, global to local, imputation guider, machine learning

1. INTRODUCTION

Missing data imputation is an important and common topic in real word, which aims to impute the uncollected and unobserved regions with reasonable values. Many imputation approaches have been proposed to handle data containing missing observations, such as multivariate time series imputation [1,2,3,4,5], image imputation [6,7,8,9], regression imputation [10,11], sentence completion [12,13,14], to cite just a few.


(a) Ground truth (b) Zero imputation (c) GAIN (d) MisGAN

(e) Missing mask (f) Mean imputation (g) GAMIN (h) Ours

Fig. 1: Some examples of challenges in missing data imputation. (a) Ground truth (GT) and (e) missing mask of which missing components are colored black. Visualization of (b) Zero imputation, (f) Mean imputation, (c) GAIN [18], (g) GAMIN [16], (d) MisGAN [17] and (h) Our proposed G2LGN.

Recently, a few significant methods about missing data imputation based on the lately prevalent generative adversarial network (GAN) [15] have been performed [16,17,18]. Most of these methods utilize generator and discriminator to learn the information of unobserved regions. Generator generates and imputes missing data to deceive the discriminator, while discriminator is used to discriminate between the imputation and fake data. Although these methods enhance the characteristic expression and follow the data distribution compared to the traditional methods, local homogenous regions especially in boundary (as shown in the blue and red boxes in Fig. 1) and the reasonable of the imputed data are still the challenging issues that negatively impacts the missing data imputation results. Essentially, there are two main reasons resulting in these issues difficult to solve. First, GAN-based methods pay more attention to make the distribution of generated data approximate the distribution of original data as a whole. Hence, the local detailed representation of the imputation is not be accurately accessed. Second, the existing methods capture the random noise to feed the model initially [15], which ignore the guiding results between different levels from global to local. Therefore, the inadequate imputation information should be effectively availed.

In this paper, we propose a novel unsupervised learning of GAN-based imputation model G2LGN to deal with missing data imputation, which consists of a Global-Impute-Net (GIN), a Local-Impute-Net (LIN) and a Impute Guider Model (IGM). GIN captures the global distribution from the entire dataset and initially generates the imputation as a whole. Considering the clutter local area and the improper results, we design IGM to stretch the information from global to local and to guide LIN to refine the local regions

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

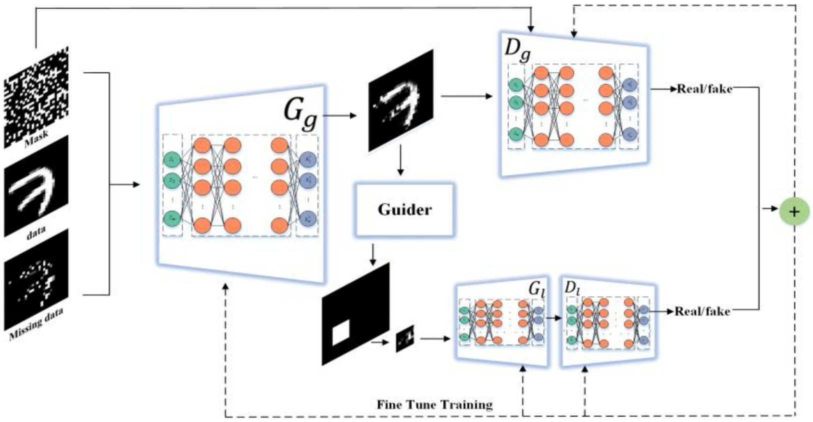


Fig. 2: The overall architecture of Global to Local Guiding Network (G2LGN).

especially the boundary of imputation results. Our G2LGN learns the guide information between global and local and refine local regions enhancing the imputation performance. Hence, the proposed G2LGN imputes more reasonable values and inadequate local regions from rough to accurate.

To sum up, the major contributions can be summarized as follows:

(1) We propose a novel G2LGN for missing data imputation. This network is designed to equip with three sub-networks. The GIN can generate imputation as a whole, while the LIN guided by IGM refines the local region especially for inadequate areas.

(2) Our method compared with 10 missing data imputation methods verify the effectiveness of our G2LGN. All the experimental results illustrate that our method outperforms on numeric datasets and image dataset. Furthermore, the ablation study demonstrates that IGM and LIN perform their effectiveness and superiority for dealing with missing data imputation.

2. PROPOSED METHODS

In this paper, we propose a generative adversarial guider imputation network (GAGIN) for missing data imputation. Our GAGIN receives a data missing completely at random and outputs the imputation using the concept of guide. Fig. 2 illustrates the overall architecture of the proposed GAGIN, which involves a Global-Impute-Net (GIN), a Local-Impute-Net (LIN) and a Impute Guider Model (IGM).

2.1 Global-Impute-Net (GIN)

We design GIN to focus on the entire missing regions to generate and impute data as a whole. In the generator, we take the missing data $\tilde{X}$, the missing mask M, and a noise variable Z as input, while outputs a vector of generated data $\bar{X}$. Obtained by taking the partial observation of $\tilde{X}$ and replacing each missing region with the corresponding value of $\bar{X}$, $\hat{X}$ corresponds to the completed data vector. Thus, we define $\bar{X}$ and $\hat{X}$ in Eqs. (1)(2) as below:

$$\bar{X} = G_g(\tilde{X}, M, (1-M) \odot Z) \tag{1}$$

$$\hat{X} = M \odot \tilde{X} + (1-M) \odot \bar{X} \tag{2}$$

where G_g is defined as a function transforming the unobserved data to generated data for every component and $\odot$ denotes element-wise multiplication.

There is a global discriminator D_g used to train our GIN model, which tries to criticize whether each component of an input is imputed or not. The missing mask M and $\hat{X}$ are combined to feed in D_g and output a value in the [0,1] range. Thus, the loss function for the adversarial global impute net is defined in Eq. (3) as follows:

$$L^g_{adv} = \sum_{i=1}^{s} L_{adv}(\, G^i_g\, ;\, D^i_g(\hat{X}, M)) \tag{3}$$

where the i-th component of D^i_g corresponds to the probability that the i-th component of $\hat{X}$ was observed.

2.2 Impute Guider Model (IGM)

For missing data imputation, it is critical for imputers to explore the felicitous structure and the appropriate local information specially in smooth homogeneous regions. However, the previous methods learn a variety of global information and treat all characteristics without distinction so that the finer details are ignored. Based on this observation, we propose an IGM to guide the LIN according to the information of the GIN result $\hat{X}$ and missing mask to meet both local refinable and values reasonable. Therefore, the global imputation results act as a prior to leading the generation and adjusting the local imputation results. Each local information is extracted from the intermediate imputation guider by the fully connected layer. To model

	Spam	Credit	letter
MICE	0.0691±0.0007	0.2574±0.0028	0.1537±0.0029
MissForest	0.0564±0.0011	0.1991±0.0005	0.1542±0.0018
Matrix	0.0604±0.0008	0.2533±0.0007	0.1448±0.0004
EM	0.0754±0.0009	0.2419±0.0026	0.1575±0.0006
KNN	0.0581±0.0022	0.1923±0.0015	0.1302±0.0007
GAIN18	0.0529±0.0011	0.1801±0.0018	0.1296±0.0005
Ours	0.0500±0.0008	0.1536±0.0010	0.1229±0.0004

Table 1: Comparison of the RMSE on UCI dataset with 20% missing rate(Average±Std of RMSE)

impute guider of intermediate results $\widehat{x_{cdd}} \in R^{local} \times R^{local}$, our proposed method IGM can be summarized as the following three steps as illustrated in Fig.3. 1) dividing the whole GIN results into few partitions and search candidate local region $f_{search}^{local}(\cdot)$ via Eq. (4); 2) digging the inter-imputation relationship ε using the extracted x_{cdd} via the multi-layer perceptron $f_{FC}(\cdot)$ in Eq. (5); 3) fusing missing mask and intermediate results via Eq. (6).

$$x_{cdd} = f_{search}^{local}(\hat{X}, M) \quad (4)$$

$$\varepsilon = f_{FC}(x_{cdd}) \quad (5)$$

$$\widehat{x_{cdd}} = \left(C(E(\varepsilon), M)\right) \quad (6)$$

where $f_{FC}(\cdot)$ is fully-connected layers with activation relu, $E(\cdot)$represents expanding the spatial dimension of ε to that of M, and $C(\cdot)$ is the concatenation operation.

2.3 Local-Impute-Net (LIN)

Benefit from the results of IGM, the proposed LIN pays more attention to partial regions relating to the missing location. The GIN takes the full data as input to recognize global consistency of the scene, while the LIN guided by IGM focuses on a small region around the inadequate imputation area in order to refine the quality of more detailed appearance. Similar to Eq. (3), the adversarial local impute net loss is defined as follows:

$$L_{adv}^{l} = \sum_{i=1}^{s} L_{adv}(\, G_l^i(\widehat{x_{cdd}}) \,;\, D_l^i(\hat{X}, M)) \quad (7)$$

Finally, the outputs of the global and the local discriminators are concatenated together into a single vector, which is then processed by a single fully-connected layer, to output a continuous value. A sigmoid transfer function [41] is used so that this value is in the [0,1] range and represents the probability that the data is real, rather than imputed.

2.4 Model Training for Imputation

We design and joint two training loss functions of the proposed G2LGN: the observation loss for training stability and the adversarial loss for improving the imputation performance.

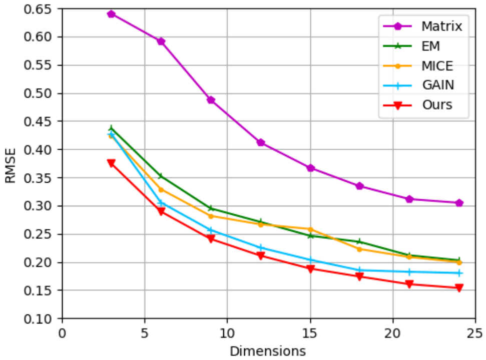


Fig. 4: Comparison of the RMSE performance by the different methods on Credit dataset for various dimension numbers.

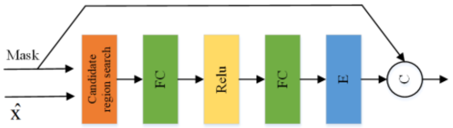


Fig. 3: The workflow of our Impute Guider Model (IGM), where *E* and *C* denote the expanding and concatenation operations, respectively.

The objective of imputation is to minimize the difference between the imputed values of the observed components and the real observed values. For GIN, the observation loss is given by:

$$L_{obs}^{g} = d(M \odot \tilde{X}, M \odot G_g(\tilde{X}, M, (1-M) \odot Z)) \quad (8)$$

where d represents distance between the imputed data and real data.

Similarly, we obtain the observation loss for the LIN and Eq. (9) shows the whole observation loss.

$$\mathcal{L}_{obs} = L_{obs}^{g} + L_{obs}^{l} \quad (9)$$

As the before mentioned GIN and the LIN are both adversarial networks, the missing data imputation is well trained between with generators and discriminators. To obtain global consistency and finer details, we define the adversarial loss as follows:

$$\mathcal{L}_{adv} = L_{adv}^{g} + L_{adv}^{l} = \sum_{i=1}^{s} L_{adv}\left(G_g^i \,;\, D_g^i(\hat{X}, M)\right) + \sum_{i=1}^{s} L_{adv}(\, G_l^i(\widehat{x_{cdd}}) \,;\, D_l^i(\hat{X}, M)) \quad (10)$$

The overall imputation losses jointed these two functions are defined as below:

$$\mathcal{L}_{imp} = \mathcal{L}_{adv} + \mathcal{L}_{obs} \quad (11)$$

3. EXPERIMENTS

3.1 Datasets and Experimentation Details

Datasets. We evaluate both on numeric datasets and image dataset for missing data imputation tasks: UCI Machine Learning Repository [28] and MNIST [29]. Like [18], we select four real-world datasets (Spam, Credit, Letter) to evaluate the imputation performance. MNIST is a dataset of handwritten digits images of size 28×28 containing 70,000 images. We use the provided 60,000 as training set, and the remaining 10,000 images as testing set.

Experimentation Details. During the training process, our G2LGN parameters is initialized by Xavier [30,37]. The dimensional vector Z is the same as the input data. Our whole network is trained by Adam optimizer [31] with learning rate 1e-3 and the batch size is set to 128. We implement our model based on TensorFlow [32] framework using python 3.7 [39] and our experiments run on a Nvidia RTX 2080Ti GPU. Due to demonstrate the performance of the proposed model fairly, as to all the compared methods, we implement with the same FC architecture that only fully-connected layers for both the generators and the discriminators.

3.2 Performance Comparison on UCI and MNIST missing dataset

Evaluation on various missing rates and dimensions for UCI missing dataset. Table 1 shows the comparison results between some other imputation methods. We investigate the influence of the comparison for various dimensions as shown in Fig.4. The RMSE performances decrease with the increasing of dimensions. The red line shows the superiority of RMSE with different dimensional datasets. We conclude that our G2LGN is also robust to the number of dimensions. Even though the RMSE of each algorithm decreases as missing rates increase, our G2LGN consistently outperforms the benchmarks owing to the proposed G2LGN can capture the global information and learn the local relationship of the unobserved values guided by IGM. Besides, our method can impute the missing data with more accurate values.

Qualitative Comparison on MNIST missing dataset. Fig.5 shows the visual imputation results generated from different methods on MNIST dataset of 50% missing rate. The (c) to (e) columns are the traditional methods, i.e. zero-imputation, mean-imputation and Matrix based imputation while the (f) to (h) columns are the GAN based methods for imputation recently as GAIN, MisGAN and GAMIN. Intuitively, the zero-imputation method and the mean-imputation method cannot produce valuable imputations. The matrix-based imputation generates the blurred images with stars. Furthermore, images imputed by GAN based methods possess the unclear and insufficient boundary in detail. It is apparent from Fig. 5 that our imputation results of the proposed method present the clearest and the most precise boundary particularly in visual performance.

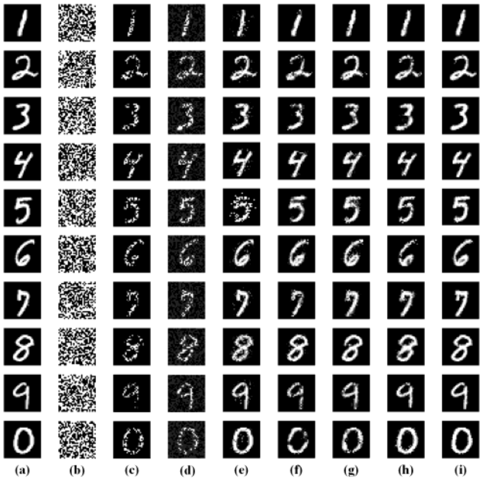


Fig. 5: Imputation results of 50% missing: (a) groundtruth, (b) missing mask of which missing components are colored black, (c) impute with zero value, (d) impute with mean value, (e) impute with Matrix algorithm, (f) GAIN based imputation, (g) MisGAN based imputation, (h) GAMIN based imputation, (i) Ours G2LGN based imputation.

Table 2. Ablation analysis of our proposed GAGIN using FID and RMSE.

Methods	Credit	MNSIT	
	RMSE	RMSE	FID
GAGIN	**0.1536**	**0.1455**	**1.2467**
Effectiveness of LIN			
w/o LIN	0.1792	0.1803	1.6247
Effectiveness of IGM			
$w/o\ \hat{X}$	0.2176	0.2293	3.4219
$w/o\ f_{search}^{local} + w/o\ f_{FC}$	0.2035	0.2177	3.1907
$w/o\ f_{search}^{local}$	0.1904	0.2075	2.8585
$w/o\ f_{FC}$	0.1848	0.1987	2.2786

3.3 Ablation Analysis

Effectiveness of Local-Impute-Net (LIN). To validate the effectiveness of LIN, we get rid of the complete LIN and then directly obtain the results of IGM and GIN concatenated together as the final outputs. The RMSE with (Row G2LGN) and without LIN (denoted as "w/o LIN") are reported in Table 2. It can be observed that G2LGN with LIN works better than that without LIN. In addition, our method without LIN gains the best performance comparing with other methods, which also highlights the effectiveness of our subnetworks (i.e. GIN and IGM).

Analysis of Impute Guider Model (IGM). To explore the importance of IGM (i.e. $\hat{X}, f_{search}^{local}$ and f_{FC}), we conduct our experiments with four different instances. Our ablation studies about the input $\hat{X}$, f_{search}^{local} and f_{FC} are shown in the Effectiveness of IGM part in Table 2. To investigate the importance of GIN's information, we first replace the input with random noise (denoted as "w/o $\hat{X}$"). To verify the effectiveness of impute guider model, we get the RMSE and FID results without f_{search}^{local} and f_{FC} (denoted as "w/o f_{search}^{local}+ w/o f_{FC} ") by replacing with the fully connected layers. As for the choice of local region, IGM is only equipped with f_{FC} (denoted as "w/ f_{FC}"). Comparing with basic model (w/o f_{search}^{local}+ w/o f_{FC}), our IGM can decrease FID up to 1.944 and reduce RMSE up to 0.0722. Hence, our $\hat{X}$, f_{search}^{local} and f_{FC} significantly improve the results.

4.CONCLUSION

In this paper, we propose a novel global to local guiding network (G2LGN) for missing data imputation. To solve the interference of local clutter and the inaccurate imputation boundary details, we design a GIN, a LIN and an IGM. After the GIN generating and imputing data on the whole, the LIN is assigned to capture and refine local details guided by the IGM. Comprehensive experiments indicate our method has the superiority of missing data imputation. Furthermore, we plan to make an additional absolute guide imputation to enhance the performance of our method.

REFERENCES

[1] Fortuin V, Baranchuk D, Rätsch G, et al. Gp-vae: Deep probabilistic time series imputation[C]//International Conference on Artificial Intelligence and Statistics. PMLR, 2020: 1651-1661.

[2] Yonghong Luo, Ying Zhang, Xiangrui Cai, and Xiaojie Yuan. EGAN: End-to-End Generative Adversarial Network for Multivariate Time Series Imputation. Twenty-Eighth International Joint Conference on Artificial Intelligence IJCAI-19, 2019.

[3] Rubanova Y, Chen R T Q, Duvenaud D. Latent odes for irregularly-sampled time series[J]. arXiv preprint arXiv:1907.03907, 2019.

[4] Liu Y, Yu R, Zheng S, et al. Naomi: Non-auto regressive multiresolution sequence imputation[J]. arXiv preprint arXiv:1901.10946, 2019.

[5] Fedus W, Goodfellow I, Dai A M. Maskgan: better text generation via filling in the_[J]. arXiv preprint arXiv:1801.07736, 2018.

[6] Lee D, Kim J, Moon W J, et al. CollaGAN: Collaborative GAN for missing image data imputation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 2487-2496.

[7] Becker P, Pandya H, Gebhardt G, et al. Recurrent kalman networks: Factorized inference in high-dimensional deep feature spaces[C]//International Conference on Machine Learning. PMLR, 2019: 544-552.

[8] Dalca A V, Bouman K L, Freeman W T, et al. Medical image imputation from image collections[J]. IEEE transactions on medical imaging, 2018, 38(2): 504-514.

[9] Lee D, Moon W J, Ye J C. Which contrast does matter? towards a deep understanding of MR contrast using collaborative GAN[J]. arXiv preprint arXiv:1905.04105, 2019.

[10] Khosravi P, Liang Y, Choi Y J, et al. What to expect of classifiers? reasoning about logistic regression with missing features[J]. arXiv preprint arXiv:1903.01620, 2019.

[11] Cortes D. Imputing missing values with unsupervised random trees[J]. arXiv preprint arXiv:1911.06646, 2019.

[12] Brown T B, Mann B, Ryder N, et al. Language models are few-shot learners[J]. arXiv preprint arXiv:2005.14165, 2020.

[13] Tran K, Bisazza A, Monz C. Recurrent memory networks for language modeling[J]. arXiv preprint arXiv:1601.01272, 2016.

[14] Zhang X, Lu L, Lapata M. Top-down tree long short-term memory networks[J]. arXiv preprint arXiv:1511.00060, 2015.

[15] Ian J. Goodfellow, Jean Pouget-Abadie, and Mehdi Mirza. Generative Adversarial Networks. Advances in Neural Information Processing Systems, 2014, 3:2672-2680.

[16] Seongwook Yoon, and Sanghoon Sull. GAMIN: Generative Adversarial Multiple Imputation Network for Highly Missing Data. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.

[17] Steven Cheng-Xian Li, Bo Jiang, and Benjamin Marlin. Misgan: Learning from incomplete data with generative adversarial networks, 2019.

[18] Jinsung Yoon, James Jordon, and Mihaela Schaar. Gain: Missing data imputation using generative adversarial nets. In International Conference on Machine Learning, pages 5675–5684, 2018.

[19] Kantardzic Mehmed. Data Mining: Concepts, Models, Methods, and Algorithms, 2011.

[20] Ian R White, Patrick Royston, and Angela M Wood. Multiple imputation using chained equations: issues and guidance for practice. Statistics in medicine, 30(4):377–399, 2011.

[21] Daniel, J, Stekhoven, and Peter Bühlmann. Missforest--non-parametric missing value imputation for mixed-type data. Bioinformatics, 28(1):112–118, 2011.

[22] Evrim Acar, Daniel M Dunlavy, and Tamara G Kolda, and Morten Mørup. Scalable tensor factorizations with missing data. In Proceedings of the 2010 SIAM international conference on data mining, pages 701–712. SIAM, 2010.

[23] Pedro J. García-Laencina, José-Luis Sancho-Gómez, Aníbal R. Figueiras-Vidal. Pattern classification with missing data: A review[J]. Neural Computing and Applications, 2010, 19(2):263-282.

[24] Andrew T Hudak, Nicholas L Crookston, Jeffrey S Evans, David E Hall, and Michael J Falkowski. Nearest neighbor imputation of species-level, plot-scale forest structure attributes from lidar data. Remote Sensing of Environment, 112(5):2232–2245, 2008.

[25] Li M, Lin J, Ding Y, et al. Gan compression: Efficient architectures for interactive conditional gans[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 5284-5294.

[26] Shen Y, Gu J, Tang X, et al. Interpreting the latent space of gans for semantic face editing[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 9243-9252.

[27] Daras G, Odena A, Zhang H, et al. Your local GAN: Designing two dimensional local attention mechanisms for generative models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 14531-14539.

[28] M. Lichman. UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml.

[29] Y. LeCun, and C. Cortes. MNIST handwritten digit database, 2010. URL http://yann.lecun.com/ exdb/mnist/.

[30] Xavier Glorot, and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. Journal of Machine Learning Research, 2010, 9:249-256.

[31] Diederik P. Kingma, and Jimmy Lei Ba. Adam: A Method for Stochastic Optimization. Computer Science, 2014.

[32] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. Tensorflow: A system for large-scale machine learning, 2016.

[33] Suhrid Balakrishnan and S. Chopra. Collaborative ranking. WSDM '12, pages 143–152. ACM, 2012.

[34] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. Neural Information Processing Systems, 2017, pp. 6626–6637.

[35] XU, Qiantong, et al. An empirical study on evaluation metrics of generative adversarial networks. arXiv preprint arXiv:1806.07755, 2018.

[36] CHAI, Tianfeng; DRAXLER, Roland R. Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. Geoscientific model development, 2014, 7.3: 1247-1250.

[37] Kumar S K. On weight initialization in deep neural networks[J]. arXiv preprint arXiv:1704.08863, 2017.

[38] Sainath T N, Vinyals O, Senior A, et al. Convolutional, long short-term memory, fully connected deep neural networks[C]//2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2015: 4580-4584.

[39] Pajankar A. Useful Unix Commands and Tools[M]//Practical Linux with Raspberry Pi OS. Apress, Berkeley, CA, 2021: 81-89.