2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 978-1-6654-6819-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/BIBM55620.2022.9995463

scSAGAN: A scRNA-seq data imputation method based on Semi-Supervised Learning and Probabilistic Latent Semantic Analysis

Zehao Xiong, Xiangtao Chen*, Jiawei Luo*, Cong Shen, Zhongyuan Xu
College of Computer Science and Electronic Engineering, Hunan University, Changsha 410083, China
E-mail: xiongzehao@hnu.edu.cn, lbcxt@hnu.edu.cn, luojiawei@hnu.edu.cn,
cshen@hnu.edu.cn, zhongyuanxu@hnu.edu.cn
* To whom correspondence should be addressed

***Abstract*—single-cell RNA-sequencing (scRNA-seq) technology can reveal cellular heterogeneity with high throughput and resolution, facilitating the profiling of single-cell transcriptomes. However, due to some experimental factors, a large number of missing values are generated in scRNA-seq data, which are called dropout events, and this phenomenon affects the downstream analysis. Imputation is an effective denoising method, but existing imputation methods still face a huge challenge: lack of interpretability. In this study, we propose single-cell Self-Attention Generative Adversarial Networks(scSAGAN), a semi-supervised imputation method for scRNA-seq data. scSAGAN mainly uses Semi-Supervised Learning (SSL) and Probabilistic Latent Semantic Analysis (PLSA), which can not only learn the potential characteristics of different types of cells but explain their imputation behavior. In clustering experiments, scSAGAN exhibits better clustering performance than all baselines on 7 datasets. Next, we interpret the imputation behavior of scSAGAN on datasets such as Alzheimer's disease and find causative genes associated with the corresponding datasets. scSAGAN is currently an open-source method, available at https://github.com/zehaoxiong123/scSAGAN.**

***Index Terms*—single-cell RNA sequencing, cell type identification, Generative adversarial network, Semi-supervised learning**

I. Introduction

The human genome project is of great significance for exploring the origin of human life [1, 2]. In recent years, scRNA-seq data has been able to better characterize the transcriptome profile at the single-cell resolution, which can provide references for cancer treatment, biological genetics, and the discovery of specific genes [3]. However, the scRNA-seq data produce more zero values than the true expression due to the influence of experimental factors, which affect many downstream analyses including cell clustering [4].

Imputation and dimensionality reduction are both effective methods to eliminate dropout events in scRNA-seq data, and previous researchers have proposed many methods to effectively reduce the noise of scRNA-seq data. The imputation method applied to scRNA-seq still faces two challenges. (1)Interpretability: Although the existing scRNA-seq data imputation method can restore the expression level of cells very well, it cannot explain the position and value of the imputation data. (2)Inference ability: Existing scRNA-seq data imputation methods are unsupervised methods, and cannot infer labels for unknown cells.

Therefore, in this paper, we propose a semi-supervised GAN-based imputation method scSAGAN to help us eliminate dropout events and restore cellular heterogeneity. scSAGAN is based on the Self-Attention GAN (SAGAN)[5], a generative model is used to generate reliable expression profiles of scRNA-seq data. The main contributions of this paper are summarized as follows:

1) We transform the SAGAN model into a semi-supervised GAN [6] that can fit the class distribution of the data.
2) To enhance the interpretability of scSAGAN, we introduce PLSA [7], a generative model that can infer associations between topics, vocabularies, and documents in natural language processing.
3) We evaluate the scSAGAN on public datasets from multiple different platforms, and the interpretability of the scSAGAN model on Alzheimer's disease(AD) datasets.

II. Method

A. Datasets

The experimental data involved in the experiments of clustering and annotating scRNA-seq data are all from [8]. These datasets consist of 7 scRNA-seq data expression profiles from different platforms and sources. Next, we mainly use Alzheimer's disease dataset published on the GEO database to verify the interpretability of scSAGAN. Details of these datasets are recorded in (**Table 1**).

In data preprocessing, scSAGAN processes the scRNA-seq data $X \in R^{M \times N}$ in three steps, where M represents the number of cells and N represents the number of genes. First, scSAGAN select 2500 highly variable genes from all sequenced genes, which are used for training and imputation. Next, scSAGAN normalizes the expression values of these genes to the range $(0, 1)$ by the scale factor. Finally, scSAGAN converts the normalized scRNA-seq data expression profile $X \in R^{M \times N^{'}}$ into images for training, where $N^{'}$ represents the number of genes after preprocessed.

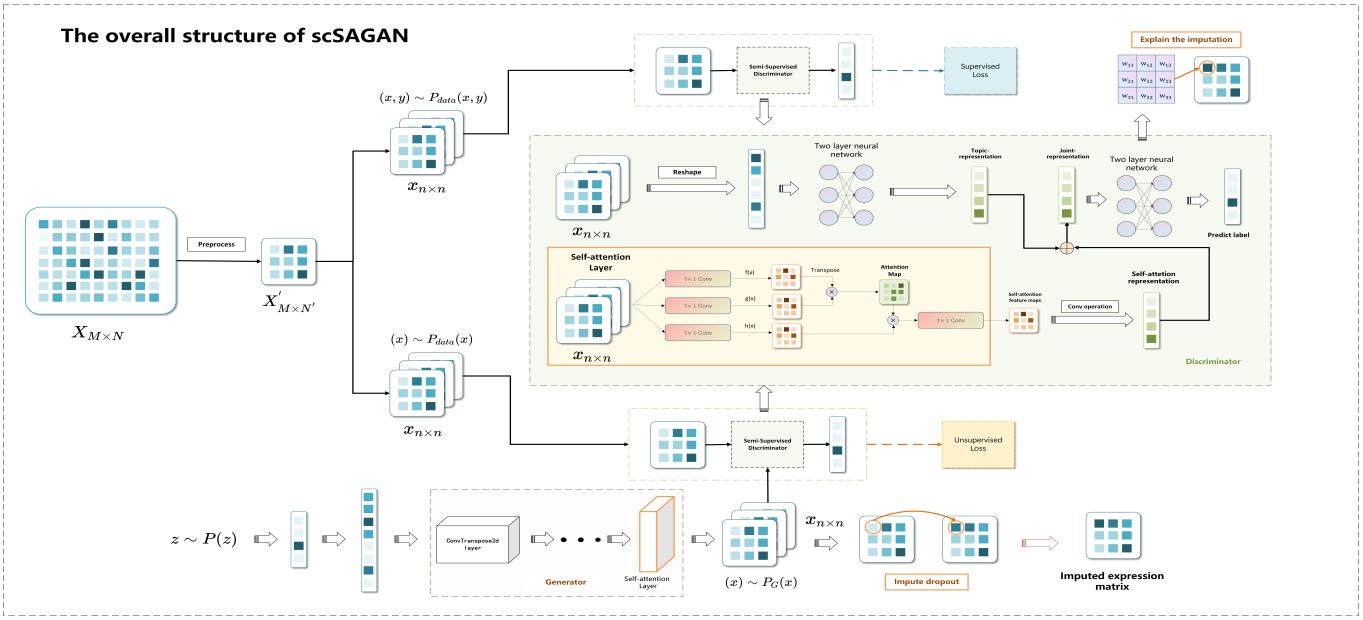

Fig. 1. The overall structure of scSAGAN. scSAGAN distinguishes labeled and unlabeled data in the preprocessed scRNA-seq data and feeds them into the model together for training. The discriminator can not only judge the true and false expression profiles of scRNA-seq data through training but also judge the cell type of the input scRNA-seq data. The generator can generate cell-specific expression profiles to impute cell-specific scRNA-seq data. The parameter matrix extracted from the neural network can be used to explain the imputation behavior.

TABLE I
SUMMARY OF THE REAL SCRNA-SEQ DATASETS

Dataset	*Cell*	*Gene*	*Class*	*Platform*
Adam	3660	23797	8	Drop-seq
Alzheimer	13215	10850	8	10x
Muraro	2122	19046	9	CEL-seq2
Qx_Bladder	2500	23341	4	10x
Qx_Spleen	9552	23341	5	10x
QS_Lung	1676	23341	6	Smart-seq2
Romanov	2881	21143	7	Smart-seq2
Young	5685	33658	11	10x

B. Overall structure of scSAGAN

We use the Self-Attention GAN [9] as the main module of scSAGAN. To better learn the characteristics of scRNA-seq data, we change the unsupervised SAGAN to a semi-supervised SAGAN, which changes the discriminator to a semi-supervised classifier and trained with a small number of labels. The overall architecture of scSAGAN is shown in **Fig. 1**.

1) Semi-Supervised Discriminator: In traditional GAN, the discriminator is generally used to determine whether the data is generated by the generator, but the discriminator of scSAGAN is used to determine which category the real cells belong to. Thus, we use $P_{model}(y = K+1|x)$ to represent the probability that the data generated by the generator is judged to be false, instead of the $1 - D(x)$ of the standard GAN.

The scSAGAN hopes that the expression values of genes in different categories of cells can replace the word frequency, and automatically learn the probability distribution of different topics in the process of training the discriminator network. The semi-supervised discriminator architecture of scSAGAN is shown in **Fig. 1**. First, we choose to transform the original data input $x \in R^{n\times n}$ into a vector $x_g \in R^{N'}$. Gene embedding $W_g \in R^{N'\times t}$ and topic embedding $W_t \in R^{t\times(K+1)}$ are set to learn the association between topics and genes, where t is the assumed number of topics. Second, we multiply x_g and W_g to get the topic representation $x_\omega \in R^t$ of the cell and get the self-attention representation $x_\tau \in R^t$ of input $x \in R^{n\times n}$ through the self-attention layer and the convolutional layer. Next, x_ω and x_τ are added by hyperparameter α to get joint representation x_α of the cell. Finally, we multiply x_α with topic embedding W_t to get a discriminant vector $x_\varrho \in R^{K+1}$ with dimension $K + 1$, which is to learn the association between topics and cell types. To support semi-supervised learning of scSAGAN, the loss is divided into three parts, supervised loss, unsupervised loss, and pseudo-sample loss. After simplification, the specific loss function is as follows:

$$\begin{aligned} L_{lable} &= -\mathbb{E}_{(x,y)\sim P_{data}(x,y)}[logP_{model}(y|x, y < K+1)] \\ &= -\mathbb{E}_{(x,y)\sim P_{data}(x,y)} log\left[\frac{exp(l_y)}{\sum_{i=1}^{K}(l_i)}\right] \\ &= -\mathbb{E}_{(x,y)\sim P_{data}(x,y)}\left\{l_y - log\sum_{i=1}^{K}(l_i)\right\} \end{aligned} \tag{1}$$

$$\begin{aligned} L_{unlable} &= -\mathbb{E}_{(x)\sim P_{data}(x)} log(1 - P_{model}(y = K+1|x)) \\ &= -\mathbb{E}_{(x)\sim P_{data}(x)} log(\frac{\sum_{k=1}^{K} exp(l_k)}{\sum_{k=1}^{K} exp(l_k) + exp(l_{k+1})}) \\ &= -\mathbb{E}_{(x)\sim P_{data}(x)} log(\sum_{k=1}^{K} exp(l_k)) + \\ &\quad \mathbb{E}_{(x)\sim P_{data}(x)} log(1 + \sum_{k=1}^{K} exp(l_k)) \end{aligned} \tag{2}$$

$$L_{fake} = -\mathbb{E}_{(x)\sim P_G(x)} log P_{model}(y = K+1|x) \\ = \mathbb{E}_{(x)\sim P_G(x)} log P_{model} log(1 + \sum_{k=1}^{K} exp(l_k)) \quad (3)$$

where the L_{lable} is to check whether the estimated labels are correct for labeled samples in the training set. The $L_{unlable}$ is to test whether the estimate is "true" for unlabeled samples in the training set. The L_{fake} is to evaluate whether the fake samples generated by the generator are estimated to be "fake". The total loss is:

$$L_D = L_{lable} + \xi(L_{unlable} + L_{fake}) \quad (4)$$

In the discriminator, we rely on the hyperparameter ξ to control the ratio between supervised loss and unsupervised loss, default $\xi = 0.3$.

2) Generator: To generate scRNA-seq data expression profiles consistent with specified categories, we design the generator of scSAGAN to be trained with specified one-hot encoding, in a similar way to CGAN [10]. We specify that the input to the generator is a simple concatenation of the vector $z \in R^t$ sampled from a Gaussian distribution and the specified one-hot encoding $z_{label} \in R^K$, resulting in $z_{can} \in R^{K+t}$. Next, z_{can} is put into the self-attention layer and the $ConvTranspose$ layer for deconvolution operation to obtain the scRNA-seq data expression profile of pseudo-cells x_p for imputation. Finally, the discriminator of scSAGAN needs to judge the category of x_p, whether it is consistent with the specified one-hot encoding z_{label}. Therefore, the loss function is defined as follows:

$$L_G = -\mathbb{E}_{(x_p,y_p)\sim P_{data}(x_p,y_p)} \left\{ l_{y_p} - log \sum_{i=1}^{K} (l_i) \right\} \quad (5)$$

where y_p is the cell label predicted by the discriminator, and the generator is trained under the guidance of the discriminator.

C. The imputation process of scSAGAN

The scSAGAN and scIGANs both use the trained generator to generate the same type of scRNA-seq data expression profile as pseudo-cell, and only impute the position where the expression value is "0". In this process, the discriminator plays the role of a "director", which specifies the position and value of the expression profile of scRNA-seq data. For labeled cells, scSAGAN imputes scRNA-seq data using the same labeled pseudo-cell expression profile. For unlabeled cells, scSAGAN first predicts the label of the cell and then imputes it.

III. Results

A. The performance of scSAGAN on downstream analysis

To evaluate the clustering performance of scSAGAN, we apply it to 7 scRNA-seq datasets and compare it with 4 state-of-the-art imputation methods. Through the verification of 7 datasets, we find that scSAGAN can achieve a better ARI with an average value of 0.55 than all baselines (**Fig. 2a, b**). Furthermore, we find that only the scSAGAN maintains the original biological characteristics and shows the separation between different cell clusters through UMAP visualization (**Fig. 2c**).

The imputation method can effectively recover missing values due to technical noise, which helps us understand the expression values produced by imputation. we conduct experiments using two gold-standard cell-annotated datasets (Zeisel [11], Klein [12]) as the benchmarks. We randomly remove 10%, 30%, and 50% of non-zero values from the gold standard scRNA-seq data to simulate the missing phenomenon caused by technical noise. The Median L1 distance and cosine similarity are used by us to measure the ability of the imputation method to restore gene expression. We find that scSAGAN achieves competitive results on the Median L1 distance and outperforms all baseline methods on the Klein dataset with an average value of 0.39. Therefore, scSAGAN can restore missing expression values in scRNA-seq data and can effectively facilitate cell clustering and visualization.

B. Predicting cell labels with scSAGAN

In the experiments, we remove 90%, 70%, and 50% of the labels in the dataset, train scSAGAN through Semi-Supervised Learning, and verify the annotation accuracy. To verify the robustness of the scSAGAN discriminator, we train the discriminator multiple times by changing the random seed when drawing training samples. scSAGAN can maintain high quality and accuracy when extracting more than 30% of training samples, but scSAGAN cannot maintain stability under 10% of training samples. The average prediction accuracy of scSAGAN on all datasets is shown in (**Fig. 3**).

C. scSAGAN explains disease-causing gene imputation

In this section, we use the PLSA model to verify the influence of important features on model training during the training process of scSAGAN. We conduct clustering experiments on real Alzheimer's disease datasets. scSAGAN imputes the AD dataset based on 8 different cell types. Since Alzheimer's disease affects gene expression in different types of cells, it is an important factor that hinders cell clustering. Among all cell types, gene expression in astrocytes, endothelial, and microglia cells is most affected. Taking astrocytes as an example, we select the highest enriched topic 46 from the 50 topics, and select the 50 highest enriched genes from it. Taking [13] as a reference, 12 of these genes are related to the differential expression of Alzheimer's disease, which proves that Alzheimer's disease is an important factor affecting cell clustering. Through the MalavCard database, we find that the remaining 17 genes are all associated with diseases such as cancer and nervous system.

IV. Discussion

To mitigate the impact of dropout events on downstream analysis of scRNA-seq data, imputation is an efficient method. However, due to the lack of inferability and interpretability, existing imputation methods are difficult to infer real scRNA-seq data. We propose scSAGAN, a Semi-Supervised Learning interpretable scRNA-seq data imputation method, which uses SAGAN and PLSA as the basic architecture to impute the

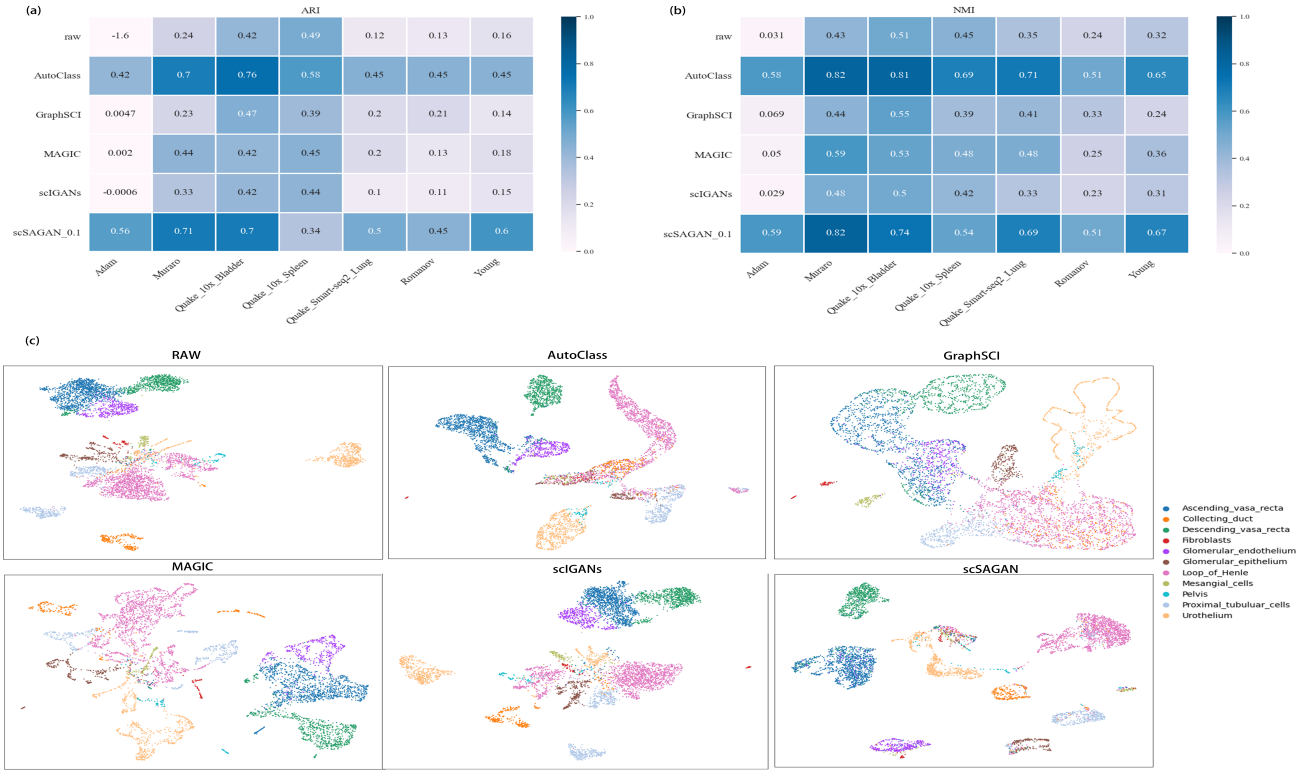

Fig. 2. Clustering results of scSAGAN. (a) Visualization of ARI metrics for multiple different alignment methods on 7 scRNA-seq datasets. (b) Visualization of NMI metrics for multiple different alignment methods on 7 scRNA-seq datasets. (c) UMAP plot for raw and imputed data on Young dataset.

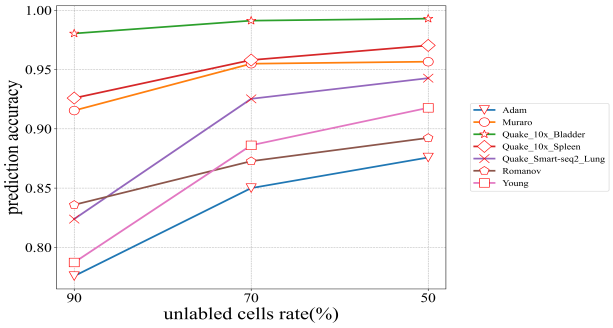

Fig. 3. The average prediction accuracy of scSAGAN on 7 different scRNA-seq datasets trained with 10%, 30%, and 50% labeled data.

scRNA-seq data dropout events. In this paper, we demonstrate the potential of semi-supervised learning in the field of scRNA-seq data imputation and provide interpretability for imputation methods, and we hope that scSAGAN can guide future scRNA-seq data imputation interpretability studies.

Acknowledgment

This work has been supported by the Nature Science Foundation of China (Grant No. 62032007, 61873089).

References

[1] Francis S Collins, Michael Morgan, and Aristides Patrinos. The human genome project: lessons from large-scale biology. *Science*, 300(5617):286–290, 2003.

[2] Francis S Collins, Ari Patrinos, Elke Jordan, Aravinda Chakravarti, Raymond Gesteland, LeRoy Walters, members of the DOE, and NIH planning groups. New goals for the us human genome project: 1998-2003. *science*, 282(5389):682–689, 1998.

[3] Laura González-Silva, Laura Quevedo, and Ignacio Varela. Tumor functional heterogeneity unraveled by scrna-seq technologies. *Trends in cancer*, 6(1):13–19, 2020.

[4] Emma Pierson and Christopher Yau. Zifa: Dimensionality reduction for zero-inflated single-cell pneumophila analysis. *Genome Biology*, 16(243), 2015.

[5] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International conference on machine learning*, pages 7354–7363. PMLR, 2019.

[6] Augustus Odena. Semi-supervised learning with generative adversarial networks. *arXiv preprint arXiv:1606.01583*, 2016.

[7] Thomas Hofmann. Probabilistic latent semantic analysis. *arXiv preprint arXiv:1301.6705*, 2013.

[8] Duc Tran, Hung Nguyen, Bang Tran, Carlo La Vecchia, Hung N Luu, and Tin Nguyen. Fast and precise single-cell data analysis using a hierarchical autoencoder. *Nature communications*, 12(1):1–10, 2021.

[9] Alex Schuurman, Tom Reijnders, Anno Saris, Ivan Ramirez-Moral, Michiel Schinkel, Justin de Brabander, Christine van Linge, Louis Vermeulen, Brendon Scicluna, Willem Wiersinga, et al. Integrated single-cell analysis unveils diverging immune features of covid-19, influenza and other community-acquired pneumonia. 2021.

[10] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[11] Amit Zeisel, Ana B Muñoz-Manchado, Simone Codeluppi, Peter Lönnerberg, Gioele La Manno, Anna Juréus, Sueli Marques, Hermany Munguba, Liqun He, Christer Betsholtz, et al. Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq. *Science*, 347(6226):1138–1142, 2015.

[12] Allon M Klein, Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A Weitz, and Marc W Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, 2015.

[13] Alexandra Grubman, Gabriel Chew, John F Ouyang, Guizhi Sun, Xin Yi Choo, Catriona McLean, Rebecca K Simmons, Sam Buckberry, Dulce B Vargas-Landin, Daniel Poppe, et al. A single-cell atlas of entorhinal cortex from individuals with alzheimer's disease reveals cell-type-specific gene expression regulation. *Nature neuroscience*, 22(12):2087–2097, 2019.