Computational and Structural Biotechnology Journal 23 (2024) 3358–3367

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research Article

AFSC: A self-supervised augmentation-free spatial clustering method based on contrastive learning for identifying spatial domains

Rui Han [a], Xu Wang [a], Xuan Wang [a,d], Yadong Wang [b,c], Junyi Li [a,c,d,*]

[a] School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen, Guangdong 518055, China
[b] Center for Bioinformatics, Faculty of Computing, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China
[c] Key Laboratory of Biological Bigdata, Ministry of Education, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China
[d] Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies, Harbin Institute of Technology (Shenzhen), Shenzhen, Guangdong 518055, China

ARTICLE INFO

Keywords:
Spatial Transcriptomics
Spatial Clustering
Self-supervised Clustering
Contrastive Learning

ABSTRACT

Recent research in spatial transcriptomics allows researchers to analyze gene expression without losing spatial information. Spatial information can assist in cell communication, identification of new cell subtypes, which provides important research methods for multiple fields such as microenvironment interactions and pathological processes of diseases. Identifying spatial domains is an important step in spatial transcriptomics analysis, and improving spatial clustering methods can benefit for identifying spatial domains. In addition to eliminating noise in original gene expression, how to use spatial information to assist clustering has also become a new problem. A variety of calculating methods have been applied to spatial clustering, including contrastive learning methods. However, existing spatial clustering methods based on contrastive learning use data augmentation to generate positive and negative pairs, which will inevitably destroy the biological meaning of the data. We propose a new self-supervised spatial clustering method based on contrastive learning, Augmentation-Free Spatial Clustering (AFSC), which integrates spatial information and gene expression to learn latent representations. We construct a contrastive learning module without negative pairs or data augmentation by designing Teacher and Student Encoder. We also design an unsupervised clustering module to make clustering and contrastive learning be trained together. Experiments on multiple spatial transcriptomics datasets at different resolutions demonstrate that our method performs well in self-supervised spatial clustering tasks. Furthermore, the learned representations can be used for various downstream tasks including visualization and trajectory inference.

1. Introduction

The different functions of biological tissues largely depend on the spatial environment in which different types of cells live, and the relative position of gene transcription and expression in tissues is crucial for analyzing their biological function and describing biological interaction networks [1]. Spatial transcriptomics technologies can not only obtain transcriptomic information of the research subject, but also preserve its spatial information, providing valuable insights for research and diagnosis. Spatial transcriptomics technologies are generally classified to two types: one is based on in situ hybridization and fluorescence microscopy spatial transcriptomics methods (including seqFISH [2,3], seqFISH+ [4], MERFISH [5,6], STARmap [7], and FISSEQ [8]) for high-resolution and accurate detection of the spatial distribution of transcripts, but they are limited in the total number of detectable RNA transcripts; the other is based on next-generation sequencing spatial transcriptomics methods, such as ST [9], Slide-seq [10], Slide-seqV2 [11], HDST [12], and 10x Visium (https://www.10xgenomics.com/), which use spatial barcodes to capture mRNA transcripts across tissue cross-sections, enabling capture of RNA expressed on the entire transcriptome scale in space, but each capture point (radius 10–100 μm), i.e. spot, contains multiple cells. These spatial transcriptomics technologies make it easier to reveal the complex transcriptome structure of tissues and the pathogenesis of disease [13,14]. An important task of spatial transcriptomics analysis is to perform unsupervised clustering [15,16], which is to assign cells or spots to spatial domains without resorting to labels. At the same time, a series of problems such as high-dimensional sparsity of gene expression data, frequent drop-out events that introduce erroneous zero values during sequencing, difficulty in effectively utilizing spatial location information, and differences in spatial

* Corresponding author at: School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen, Guangdong 518055, China.
E-mail address: lijunyi@hit.edu.cn (J. Li).

https://doi.org/10.1016/j.csbj.2024.09.005
Received 22 May 2024; Received in revised form 7 September 2024; Accepted 7 September 2024
Available online 10 September 2024
2001-0370/© 2024 The Author(s). Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

transcriptomics data caused by differences in sequencing technologies make unsupervised clustering analysis methods for spatial transcriptomics more challenging and meaningful [17–19].

Currently, many research methods have considered incorporating spatial information into clustering analysis and have developed different methods for identifying spacial domains. BayesSpace and Giotto use statistical method to increase the resolution of spot-resolution spatial transcriptomics data through spatial neighbors [20,21]. However, these probabilistic-based models cannot effectively learn the latent representations of cells or spots. Therefore, methods based on deep learning have been introduced. SEDR [22] learns latent gene representations through Auto-Encoder and reconstructs the spatial structure through Variational Graph Auto-Encoder. SpaGCN [23] uses histopathological images and spatial locations to build relationships between spots and uses graph convolutional network (GCN) [24] to learn latent representations. STAGATE [25] combines cell location and gene expression to learn node representations through Graph Attention Auto-Encoder, adaptively learns the similarity between adjacent nodes. CCST [26] uses the DGI [27] model to maximize mutual information and uses the GCN encoder to aggregate features of adjacent nodes. As deep learning develops, contrastive learning methods have attracted researchers' attention due to their ability to extract advanced semantic features from data. Contrastive learning is a discriminative learning method in which similar samples learn similar representations, while dissimilar samples learn representations away from each other. The initial design of contrastive learning was to train a representation extraction model by constructing positive pairs and negative pairs[28]. For each sample, the positive samples are obtained by this sample through data augmentation, and the negative samples are the other samples. These models aim to learn representations by maximizing the similarity between positive pairs and minimizing the similarity between negative pairs. ConST [29] extracts the morphological features of each spot from histopathological images through pre-trained Masked Autoencoder model [30], and then integrates gene expression and pre-extracted morphological features to feature vectors, which are input to GCN to learn latent representations. GraphST [31] designs a contrastive learning strategy that captures both the local and global context of nodes. ConGI [32] also introduces histopathological images and designs three contrastive loss functions. However, existing methods use data augmentation to construct positive pairs and negative pairs, and commonly used data augmentation methods for graph structures and gene expression matrices, such as adding and removing edges, randomly masking and shuffling, inevitably destroy the biological meaning contained in the original structure which makes positive pairs can not be constructed effectively. Spatial transcriptomics has great prospects for development [33–35], and technical methods are constantly being updated on various platforms. Differences in technology across platforms can result in differences in data, while the ability of existing methods to be compatible with data from different platforms is not outstanding.

Based on the issues mentioned above, we propose a self-supervised contrastive learning model AFSC for spatial clustering. Existing methods all follow the common contrastive learning process based on data augmentation, which ignore the particularity of spatial transcriptomics data. Our method without data augmentation can effectively improves the shortcomings of existing methods in this regard, which can more play the advantages of contrastive learning in spatial clustering tasks. By integrating spatial positions and gene expression, we learn latent representations using a contrastive learning model with a specific positive pairs selection strategy. We introduce a clustering loss in the unsupervised clustering module to guide the training process, along with the contrastive loss, which makes the learned latent representations can be optimized for clustering tasks. In order to demonstrate the performance of our method, we conducted multiple experiments on seven datasets at spot and single-cell resolution. We've selected evaluation metrics to evaluate from different perspectives. Specifically, we use Adjusted Rand index (ARI) [45], Normalized Mutual Information (NMI) [46], Fowlkes-Mallows index (FMI) [47] and Jaccard index [48,54] to evaluate datasets with manual annotation, as well as Silhouette Coefficient (SC) [49] and Davies-Bouldin index (DBI) [50], to evaluate datasets without manual annotation. We also conducted experiments on runtime and max RAM usage of GPU on two datasets as shown as in the Supplementary Material Table S1. Experiments and comparative analysis on multiple datasets indicate that our method performs well in identifying spatial domains, and overall, our method is more universal for different datasets. Moreover, the representations learned from our model can also be applied to downstream tasks such as visualization and trajectory inference.

2. Datasets and materials

2.1. Dataset description

To evaluate and test the performance of our model, we collect seven datasets from multiple public platforms, including 12 slices of the human dorsolateral prefrontal cortex (DLPFC) acquired with 10x Visium (https://www.10xgenomics.com/), the breast invasive carcinoma (BRCA) acquired with 10x Visium, the anterior of the mouse brain tissues (MBA) acquired with 10x Visium, the mouse olfactory bulb (MOB) dataset from Stereo-seq [41], and the mouse hypothalamic preoptic area acquired with MERFISH [6], the mouse cortex subventricular zone (cortex_SVZ) and mouse olfactory bulb (OB) independent tissues acquired with seqFISH [51]. We choose datasets covering different resolutions, large variations in gene numbers, normal and disease tissues. We treat the processing of cells and spots equally for the two different resolution datasets, which also shows the excellent versatility and flexibility of our model.

2.2. Dataset preprocessing

From the obtained spatial transcriptomics data, we extracted the required information, including gene expression and spatial coordinates. First, a raw gene expression matrix is constructed based on gene expression, with cells (spots) and gene expression being the row and column, respectively. The raw gene expression matrix contains a lot of noise, and therefore data preprocessing is necessary. To deal with the high-dimensional sparsity of gene expression, we first remove genes that were expressed in less than three cells (spots) and then normalize and standardize the data with Scanpy package [36]. Finally, principal component analysis (PCA) is used to obtain the feature matrix X. Since our model is based on the graph constructed by the cells (spots) for graph self-supervised clustering, a graph that can better exhibit the relationships between the cells (spots) should be constructed based on spatial locations. We treat the cells (spots) as nodes, and used spatial coordinates to calculate the distances between all cells (spots). We add edges to each cell (spot) with its k nearest neighbors, and obtain an adjacency matrix. Then, we use all the distances between each node and its k nearest neighbors to get the distance distribution (Fig. 1a). The threshold is set up based on the distribution to make sure some farther neighbors won't be selected. We weight the edges based on the distances to obtain a weighted adjacency matrix A. Finally, the feature matrix X is treated as the attribute of nodes, and we obtain a weighted adjacency graph, where nodes represent all cells (spots), and edges represent neighbour relationships between all cells (spots).

We set the dimension of PCA as 100 for MERFISH dataset and 1000 for the other datasets since only 160 genes were captured in MERFISH. We set k $=$ 10for k nearest neighbors as default. The median of distance distribution is set to be the default of threshold. All these parameters can be easily adjusted accordingly to meet the requirements of specific datasets.

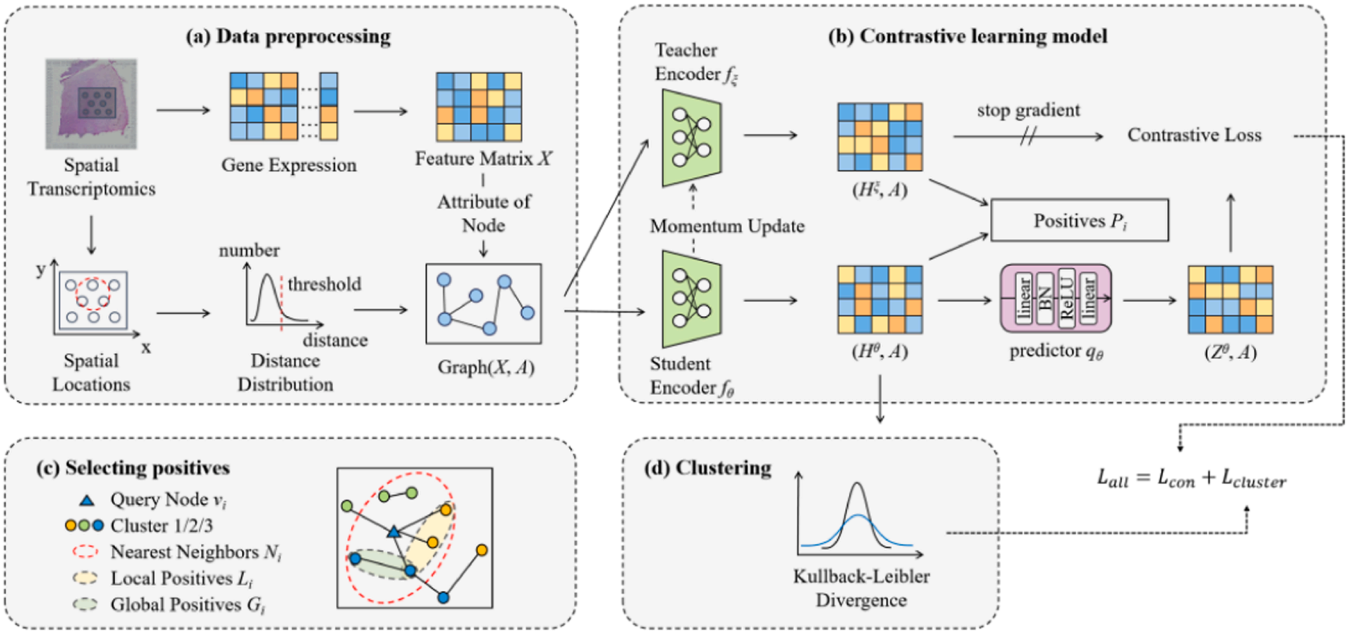

Fig. 1. A The overall architecture of AFSC. This figure shows how we obtain a weighted adjacency graph $G(X, A) = (V, E)$ from spatial transcriptomics using raw gene expression and spatial locations. The feature matrix X is obtained after preprocessing the raw gene expression and the weighted adjacency matrix A is obtained based on the distances between cells(spots) to describe their spatially relations. (b) Contrastive learning model. This figure shows the contrastive learning model based on a Siamese network. f_ξ and f_θ are two encoders of this network, and q_θ is a predictor. H^ξ and H^θ are used to design the positive set P_i of any node v_i, and Z^θ and H^ξ are used to calculate contrastive loss L_{con}. (c) Strategy for selecting positives sets. This figure shows nodes distribution on representation space, and existing edges show node relationships on the adjacency graph G. (d) Unsupervised clustering. This figure shows the introduced clustering loss $L_{cluster}$ is the Kullback-Leibler Divergence of two distributions. High-resolution images are available for viewing at https://github.com/bioszhr/AFSC/tree/main/results/figures.

3. Model and implementation

3.1. Overview of AFSC

Our model is based on contrastive learning, using a Siamese network to learn embeddings, selecting the positives sets of nodes to define the contrastive loss [37], and introducing the clustering loss to construct the total loss function. As shown in Fig. 1, after the preprocessing described above, we obtain a weighted adjacency matrix A and feature matrix X, which are respectively input into the two encoders f_ξ and f_θ of the Siamese network. The two encoders learn the node embedding H^ξ and H^θ, respectively, and the learned embeddings along with spatial locations are used to select the positives set P_i of each node. Then, H^θ is passed through the predictor to obtain Z^θ, and the contrastive loss L_{con} is designed by Z^θ and H^ξ. The loss function is to close the cosine distance between each node and its positives, achieving the goal of contrastive learning without data augmentation or negative pairs. At the same time, H^θ is used to calculate the clustering loss $L_{cluster}$ and to perform clustering. The overall loss is calculated to guide the training, and the final clustering result is obtained after training. The learned embedding H^θ can be applied to other downstream tasks.

4. Contrastive learning model

Siamese network learning latent representations.

As mentioned in the Introduction, data augmentation used by current contrastive learning methods will destroy the biological meaning contained in the original graph structure and gene expression matrix, such as adding and removing edges, randomly masking and shuffling gene expression matrix. Gene expression feature of cells are significant for distinguishing cell types, and due to the technical limitations, there are inevitably noises in gene expression data. In this case, data augmentation methods of gene expression matrix can seriously damage the feature of cells, which makes it impossible to construct high-quality positive samples. Therefore, we design a contrastive learning method that does not require data augmentation or negative pairs, which means each node only need to learn from its positives. In traditional contrastive learning frameworks, negative examples are the driving force behind model learning, because if the objective function only pulls the features of the sample and its positive samples closer, the model will easily collapse and the features of all samples will become the same. BYOL [44] was able to successfully design a contrastive learning model without negative samples for Computer Vision, which shows that it is possible to discard negative samples if positive samples can be similar yet diverse with each sample. Inspired by this, we use a Siamese network to make sure that each node can learn similar yet diverse representations with their positives.

Siamese network refers to two networks with the same structure, but their parameters are different. In this case, teacher encoder f_ξ and student encoder f_θ. f_ξ and f_θ are both one-layer GCN with the same structure and randomly initialized separately. The parameters θ of f_θ are updated via gradient descent, while the parameters ξ of f_ξ are updated by θ via a momentum update coefficient τ as follows:

$$\xi \leftarrow \tau\xi + (1-\tau)\theta \tag{1}$$

The update of ξ is slower and smoother, which effectively prevents the collapse phenomenon that may occur during the learning process, where the model cannot learn meaningful representations. After inputting the weighted adjacency matrix A and feature matrix Xobtained from preprocessing, the two encoders f_ξ and f_θ respectively learn node representations H^ξ and H^θ as follows:

$$H^\xi = f_\xi(X, A) \tag{2}$$

$$H^\theta = f_\theta(X, A) \tag{3}$$

The i-th row of H^ξ and H^θ, i.e., h_i^ξ and h_i^θ, are the different embeddings of node $v_i \in V$ learned by different encoders. For any node v_i, its positives set P_i is determined by h_i^ξ and h_i^θ. q_θ is a predictor defined as a

multi-layer perceptron (MLP) with batch normalization, and Z^{θ} is obtained by applying the predictor to H^{θ}, where z_i^{θ} is the predicted value of h_i^{θ}, as follows:

$$Z^{\theta} = q_{\theta}(H^{\theta}) \tag{4}$$

The difference between $Z^{\theta}$and H^{ξ}can be further enlarged by the feature transformation through the predictor. Considering that the difference between Z^{θ} and H^{ξ} generated by the teacher encoder f_{ξ} above has been further enlarged. At this point, for node $v_i \in V$, the two features z_i^{θ} and h_i^{ξ} have been similar yet diverse. Based on this, we define the contrastive loss function L_{con} as Eq. 5, aiming to minimize the cosine distance between each node and its positives, i.e., to minimize the cosine distance between z_i^{θ} and h_j^{ξ} as much as possible, where v_j is any node in P_i.

$$L_{con} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{v_j \in P_i}\frac{z_i^{\theta} h_j^{\xi T}}{\left\|z_i^{\theta}\right\|\left\|h_j^{\xi}\right\|} \tag{5}$$

Strategy for selecting positives set.

The model above can only ensure that the samples can learn similar yet diverse features from their positive samples, the strategy for selecting positives set is crucial for contrastive learning. For a query node v_i, the cosine distance between v_i and all other nodes, i.e., the cosine similarity between h_i^{θ} and h_j^{ξ} is calculated:

$$\mathrm{sim}(v_i, v_j) = \frac{h_i^{\theta} \bullet h_j^{\xi}}{\left\|h_i^{\theta}\right\|\left\|h_j^{\xi}\right\|}, \quad \forall v_j \in V \tag{6}$$

Given the similarity information, we calculate the k-nearest neighbors set N_i of each node v_i. Nodes in N_i are adjacent to v_i in the representation space, making the set N_i a reasonable choice for positives set of v_i. However, just considering the nearest neighbors in the representation space may not only ignore the local semantics of the graph which means neighbors of nodes in the adjacency graph, but also ignore the global semantics of the graph which means the possible clustering results. Based on this, we design Local Positives L_i and Global Positives G_i to capture the positives set of node v_i in local and global semantic contexts, respectively (Fig. 1c).

$$L_i = N_i \bigcap A_i \tag{7}$$

where A_i refers to the neighbors of v_i in the adjacency graph G. L_i considers local semantic context between nodes based on N_i.

$$G_i = N_i \bigcap C_i \tag{8}$$

where C_i refers to the nodes that are clustered by K-means and are in the same class as v_i. C_i considers global semantic context based on N_i. Finally, the positives set P_i of node v_i is designated as follows:

$$P_i = L_i \bigcup G_i \tag{9}$$

4.1. Unsupervised clustering loss

Since the graph clustering task is unsupervised, it is impossible to provide feedback during training on whether the learned representations has been optimized well. In order to make the generated node embedding better serve the clustering task, our model incorporates clustering into the training based on DAEGC [38], optimizing the training of the encoder through clustering loss. Using the embedding H^{θ} as Eq.3, k-means is first applied to initialize the clustering. This algorithm generates several clusters and obtains centroid of each cluster, and the embedding of the centroid of cluster u denotes as μ_u. One way to solve unsupervised learning tasks is to generate "soft" labels and then use these labels to supervise training. Here, soft clustering distribution q_{iu} and auxiliary distribution p_{iu} are defined, and the KL divergence is used to narrow the distance between the two distributions, simultaneously optimizing clustering and embedding:

$$L_{cluster} = KL(P \parallel Q) = \sum_i \sum_u p_{iu} \log \frac{p_{iu}}{q_{iu}} \tag{10}$$

Firstly, based on the t-distribution, we use q_{iu} to measure the similarity between the embedding z_i of node v_i and the centroid μ_u of a cluster. Then, we can calculate the probability that node v_i belongs to cluster u, which can be regarded as the soft clustering distribution of each node:

$$q_{iu} = \frac{\left(1 + \|z_i - \mu_u\|^2\right)^{-1}}{\sum_k \left(1 + \|z_i - \mu_k\|^2\right)^{-1}} \tag{11}$$

Furthermore, we need to force each node to be closer to its corresponding cluster centroid, i.e., minimize intra-cluster distance and maximize inter-cluster distance. Therefore, we define the auxiliary distribution p_{iu} to refine clustering:

$$p_{iu} = \frac{q_{iu}^2 / \sum_i q_{iu}}{\sum_k \left(q_{ik}^2 / \sum_i q_{ik}\right)} \tag{12}$$

In the auxiliary distribution, squaring can achieve an "emphasis" effect to highlight the role of high-probability "confident" nodes. During the training, the distribution actually acts as labels. Finally, we use Eq.10 to fit the difference between the two probability distributions to reach the goal of unsupervised clustering. Eq.10 also guides the entire training as the clustering loss $L_{cluster}$.

4.2. The overall goal of the model

The embedding H^{θ} and the contrastive loss L_{con} are obtained in the contrastive learning module, where H^{θ} is used in the unsupervised clustering module to obtain the clustering loss $L_{cluster}$, which together with L_{con} form the total loss function L_{all} as follows:

$$L_{all} = L_{con} + L_{cluster} \tag{13}$$

After training is done, based on the soft clustering distribution q_{iu}, we obtain the estimated label s_i for each node, which is the clustering result:

$$s_i = \arg\max_u q_{iu} \tag{13}$$

5. Experiment results

5.1. Experimental setups

5.1.1. Evaluation criteria

To validate the performance of clustering, we use internal metrics, ARI, NMI and FMI, to measure the similarity between clustering results and manual annotation, as well as external metrics, SC and DBI, to measure the overall clustering performance without manual annotation. The brief conceptual explanations of each metric are showed in Supplemental Material.

5.1.2. Baseline methods

Several baseline methods were selected to evaluate our model on different datasets, including Seurat [39], Louvain, stLearn [52], SEDR, SpaGCN, Giotto, BayesSpace, Stardust [53], CCST, conST, conGI, GraphST and ConSpaS [42]. These methods cover probability models based on early machine learning, graph neural network models based on deep learning, and recent models based on contrastive learning. As for Seurat, stLearn, SEDR, SpaGCN, Giotto, BayesSpace, Stardust, CCST, conST, ConSpaS, we use default parameters and settings in the whole program. As for Louvain, we use our preprocessing method and then use Louvain in clustering. As for conGI and GraphST, we use Louvain

algorithm [40] in clustering and default parameters and settings in other parts of the program.

6. Main results

6.1. Application to spatial clustering at spot resolution

We first evaluate spatial clustering performance of our model on the spatialLIBD human dorsolateral prefrontal cortex (DLPFC) (https://www.10xgenomics.com/). This dataset includes 12 tissue slices, each depicting four or six neuronal layers and white matter (WM). The number of spots in each slice ranged from 3460 to 4789, with 33,538 genes captured. Boxplots were created for ARI, NMI, and FMI (Fig. 2a) to compare results of the 12 slices. AFSC achieve the highest median ARI and FMI values and the second highest median NMI value. We also visualize the results of tissue slice 151508 (Fig. 2b & Supplemental Figure S1), and CCST, conGI, and AFSC restore a clear tissue structure matching the structure of manual annotation.

Next, we analyze the breast invasive carcinoma (BRCA) and the anterior of the mouse brain tissues (MBA) from 10x Visium (https://www.10xgenomics.com/). The BRCA dataset consists of 20 regions and the number of spots is 3798, with 36,601 genes captured.

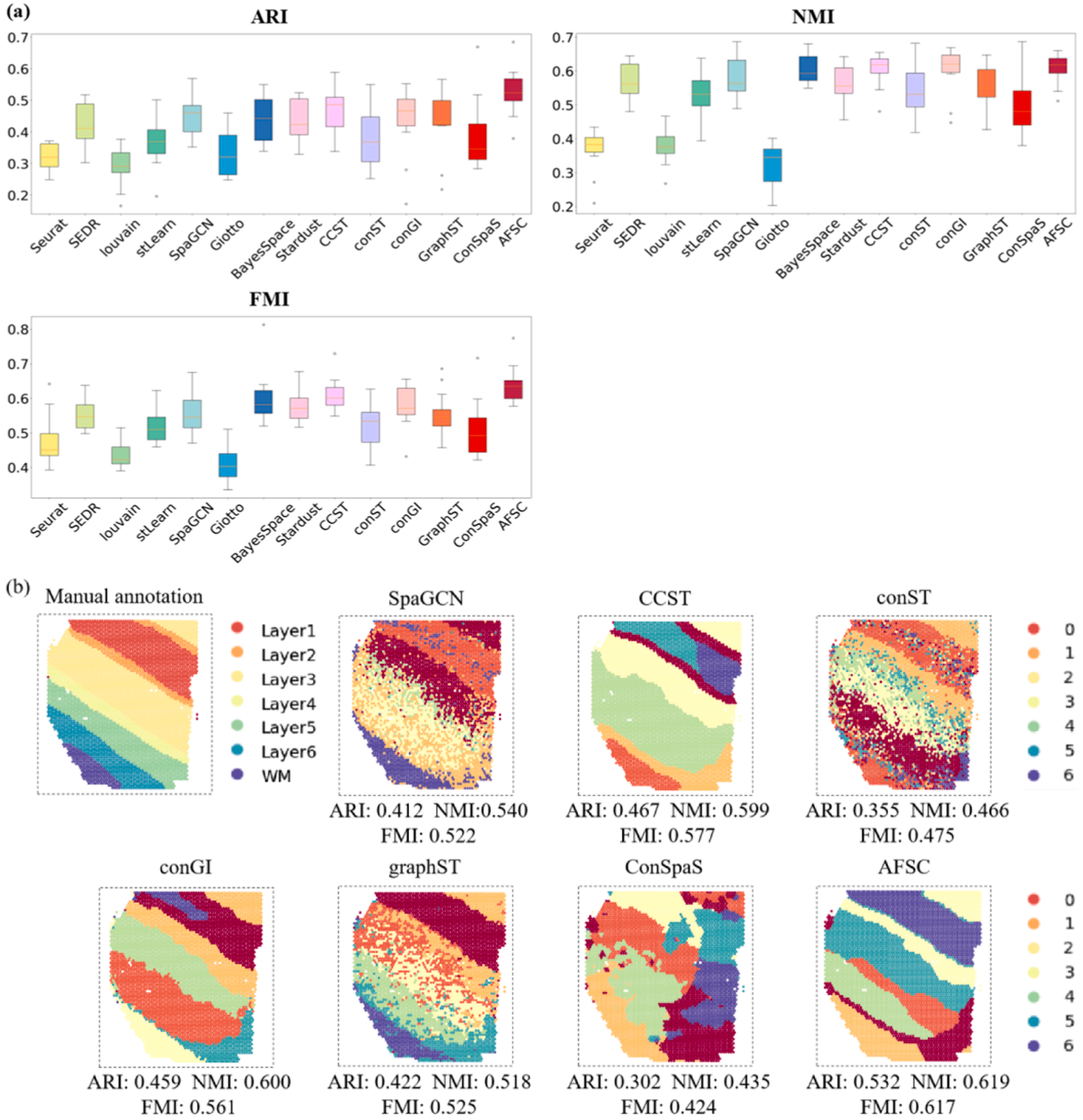

Fig. 2. Spatial domains detected in datasets at spot resolution. (a) Boxplots of ARI, NMI and FMI scores of methods applied to the 12 DLPFC slices. (b)Visualization results of manual annotation and methods applied to the tissue slice 151508 of DLPFC dataset. High-resolution images are available for viewing at https://github.com/bioszhr/AFSC/tree/main/results/figures.

The MBA dataset consists of 52 regions and the number of spots is 2695, with 32,285 genes captured.

On the BRCA dataset, AFSC achieve the highest ARI, NMI and FMI values. We also conduct visualization experiments on this dataset (Fig. 3a & Supplemental Figure S2) and provide the visualization results of the manual annotation for comparison. The visualization results show AFSC can restore clear organizational structures. Taking both clustering metrics and visualization results into consideration, AFSC outperforms other baselines on this dataset. On the MBA dataset, all methods perform bad (Fig. 3b & Supplemental Figure S3), perhaps due to the complex structure of this dataset, making it challenging to learn effective representations. Nevertheless, AFSC achieve the highest ARI and FMI values and the second highest NMI value. Overall, AFSC is proved to be more effective on these datasets at spot resolution.

In addition to these datasets, we also conduct experiments on MOB dataset [41] from Stereo-seq without manual annotation. This dataset consists of 7 regions and the number of spots is 19,109, with 14,376 genes captured. AFSC achieve the highest SC value and the best DBI value (Fig. 4), and the visualization results show that SEDR and AFSC can restore structure of mouse olfactory bulb tissue.

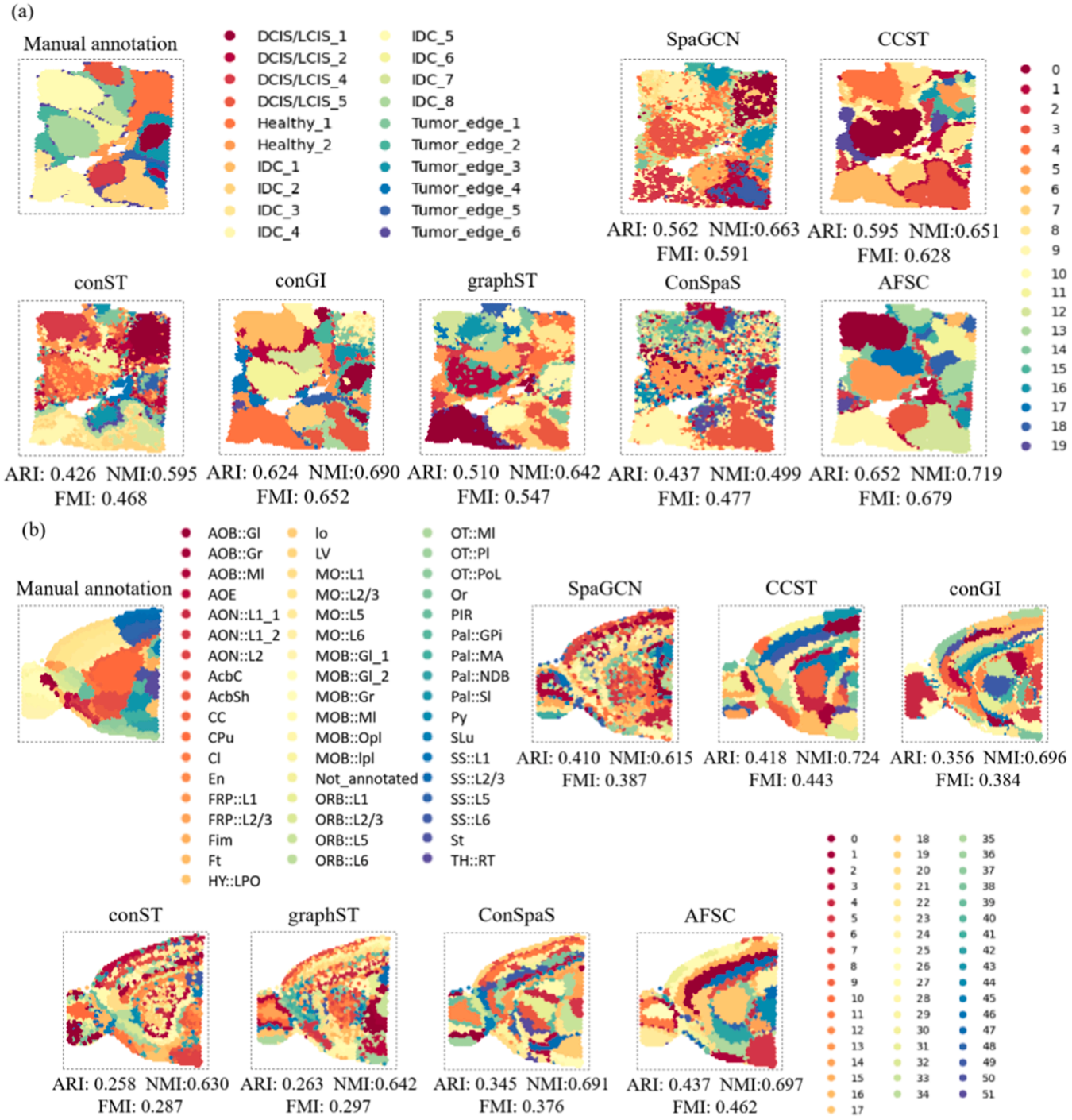

Fig. 3. Visualization results of datasets at spot resolution. (a) Visualization results of manual annotation and methods applied to the BRCA dataset. (b) Visualization results of manual annotation and methods applied to the MBA dataset. High-resolution images are available for viewing at https://github.com/bioszhr/AFSC/tree/main/results/figures.

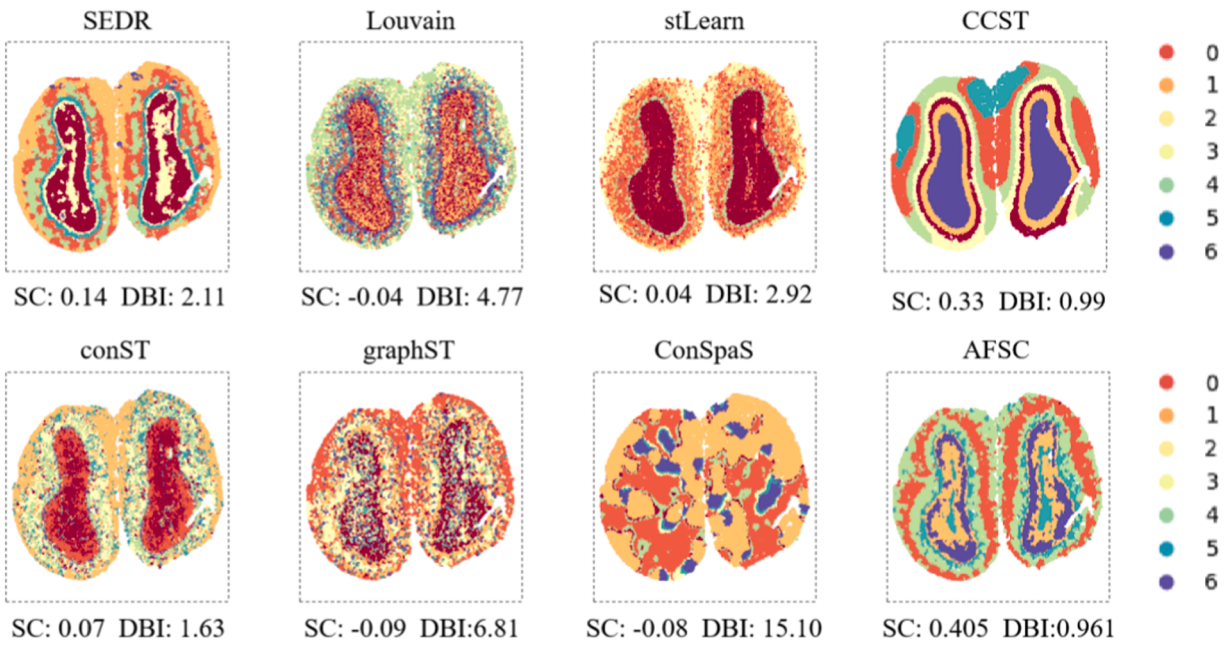

Fig. 4. Visualization results of MOB datasets. Visualization results of methods applied to MOB dataset. High-resolution images are available for viewing at https://github.com/bioszhr/AFSC/tree/main/results/figures.

Application to spatial clustering at single-cell resolution.

In order to validate the versatility of AFSC, experiments are also conducted on the datasets that reach single-cell resolution. The MERFISH dataset [6] includes multiple slices, and we choose animal No.18 and bregma= 0.11, which is divided into 15 regions. The number of cells is 4975, with 160 genes captured. The seqFISH dataset [50] includes the mouse cortex subventricular zone (cortex_SVZ), which is divided into 26 regions, and the mouse olfactory bulb (OB) independent tissues, which is divided into 25 regions. The cell number of cortex_SVZ dataset is 913 and the cell number of OB dataset is 2050, both with 10,000 genes captured.

On the MERFISH dataset, due to technical limitations, expression counts for only 160 genes were collected, most models could not learn effective representations. On the MERFISH and two seqFISH datasets, AFSC achieves the second best SC value and DBI value (Fig. 5). Although CCST achieves the highest value, it is obvious that CCST couldn't restore the structure of mouse hypothalamic preoptic region from visualization results of MERFISH dataset (Fig. 5a). Consider metrics and visualization results together, AFSC performs well on dataset at single-cell datasets which indicates that AFSC has good universality and generalization compared to other methods, and will not be greatly affected by the technology. This demonstrates that our handling of spatial positions and gene expression is reasonable, and even in datasets with extremely few genes and unclear tissue structure, we can still learn effective representations without being greatly affected by data noise.

Latent representations learned from AFSC being applied to downstream tasks.

Our model is based on deep learning, where the module for learning latent representations is a critical component. We aim to learn effective latent representations from spatial locations and raw gene expression, which not only affect clustering performance but also greatly impact other downstream tasks.

To demonstrate that the latent representations learned by our model can be effectively applied to different downstream tasks, we visualize the learned high-dimensional representations using uniform manifold approximation and projection (UMAP) [43] and construct trajectory inference graphs using the partition-based graph abstraction (PAGA) algorithm in the Scanpy package. We compare our model with baselines based on deep learning. The latent representations learned from our model are obtained after the whole training progress, including the Siamese network and unsupervised clustering.

We first perform UMAP visualization and PAGA trajectory inference on the tissue slice 151676 of DLPFC dataset, as shown in Fig. 6. We also conduct experiments on the mouse olfactory bulb (OB) independent tissues at single-cell resolution, as shown in Supplemental Figure S4. This dataset has more regions and a more complex tissue structure than the DLPFC dataset. Compared with the UMAP visualizations of the other methods, AFSC can clearly distinguish most of the spots, while some methods exhibit mixed spots from different clusters. We further generate PAGA graphs on tissue slice 151676, which displays clear layer structure of human dorsolateral prefrontal cortex. CCST and AFSC generate clear PAGA graphs that show the developmental trajectories between different layers with some correlation between adjacent layers. Results of these experiments indicate that our model can learn effective latent representations on different types of datasets.

6.2. *Ablation study*

We design a set of ablation study on the tissue slice 151676 of the DLPFC dataset (Table 1). $\tau = 0$ sets the momentum update coefficient of the Siamese network to 0, i.e., keeping parameters of the teacher encoder and the student encoder identical during training. $\tau = 1$ sets the momentum update coefficient of the Siamese network to 1, i.e., keeping parameters of the teacher encoder frozen after random initialization. w/ SGD allows parameters of the teacher encoder in the Siamese network to be updated with gradient descent. random_pos randomly selects k positives for each node. knn_pos selects the k-nearest neighbors on the representation space of each node as positives. w/o L_cluster removes the clustering loss and only uses the contrastive loss to train the contrastive learning model, and the trained H^{θ} is directly used for K-means clustering.

The experimental results show that all incomplete models have significantly decreased in three indicators, indicating that each component of our model plays an indispensable role, some of them are considerable for our model. The first three experiments are all related to the Siamese network model, indicating that if the representations learned by teacher encoder and student encoder are completely the same, the result of contrastive learning will be poor, and if the difference

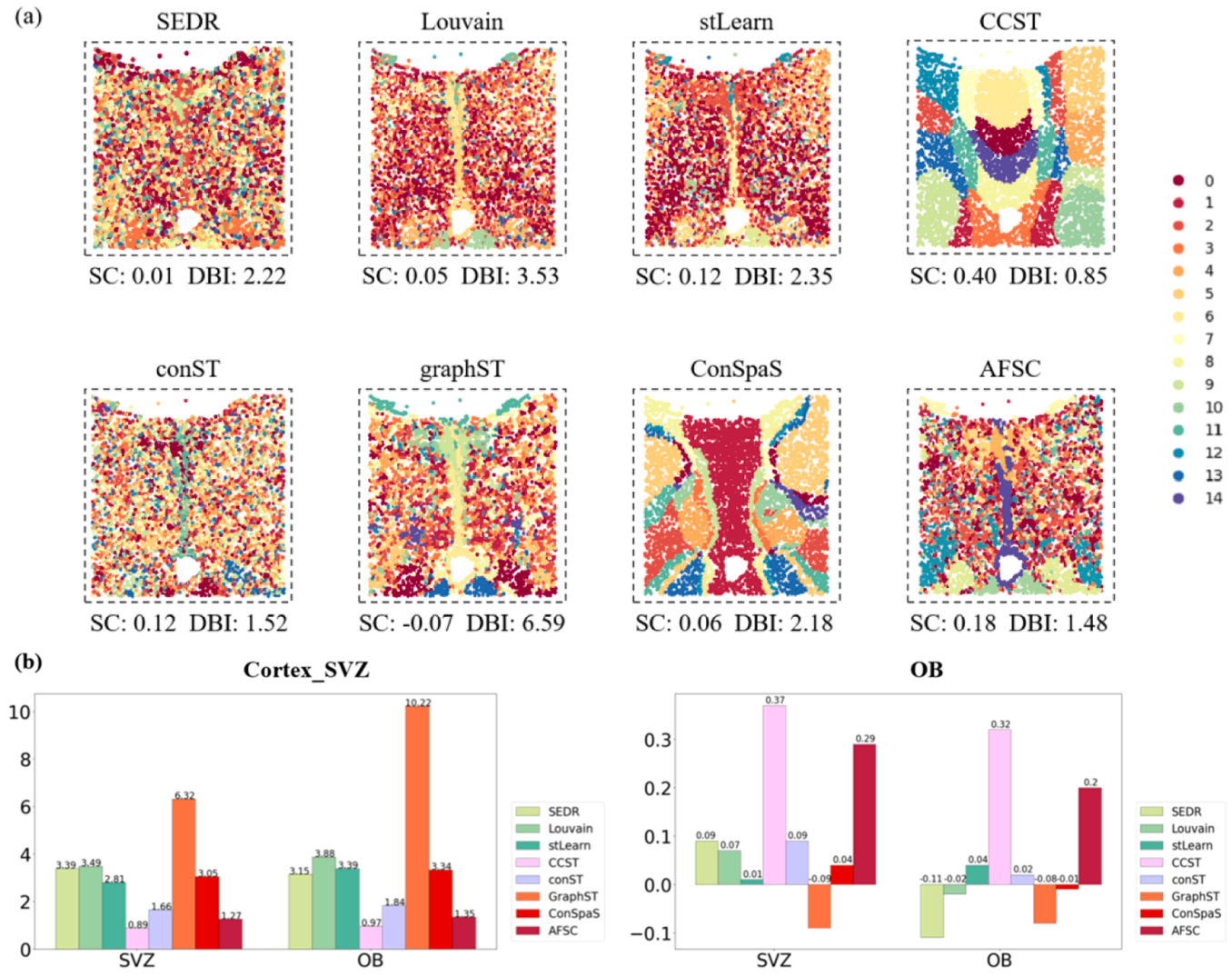

Fig. 5. Visualization results of datasets at single-cell resolution. (a) Visualization results of methods applied to the MERFISH dataset. (b) Histograms results of SC and DBI scores of Cortex_SVZ and OB datasets. High-resolution images are available for viewing at https://github.com/bioszhr/AFSC/tree/main/results/figures.

of representations is too large, it will also affect the model's performance. Evidently, our model can make the difference between the representations of nodes and their positives just right to maximize the advantages of contrastive learning. random_pos decrease the most among all models, indicating that the positives sets are crucial in our contrastive learning design. Using contrastive loss based on random positives can result in collapse phenomenon during training. knn_pos is the model with the smallest decrease in performance, indicating that the neighbors of nodes in the representation space can be used as positives sets, although there is some noise in it. Our strategy for designating positives sets can effectively eliminate these redundant nodes and improve the performance of the model. w/o L_cluster indicates that introducing clustering loss into training can improve clustering. Clustering-oriented training will be more effective than training with only contrastive loss.

7. Conclusion and discussion

Unsupervised clustering methods for identifying accurate spatial domains is critical for researching organizational functions of organisms. In this paper, we propose a new self-supervised spatial clustering method AFSC, which integrates spatial information and gene expression to learn latent representations without negative pairs or data augmentation. We also introduce a clustering loss, which allows the learned latent representations to be optimized for the clustering task, to guide training together with the contrastive loss. Multiple experiments demonstrate the versatility and validity of our method on different datasets in the task of self-supervised spatial clustering. Moreover, the finally learned representations from the model can also be used in other downstream tasks, such as visualization and trajectory inference.

Despite of good performance of AFSC, there is still room for improvement. Since many datasets do not provide histopathological images, we did not introduce images to ensure the versatility of the model. While existing studies have demonstrated that histopathological images are effective for spatial clustering, and as technology continues to evolve, it is likely that more datasets will provide images. Our work has been shown to outperform some methods using histopathological images, and in subsequent work, we will introduce histopathological images to further improve our model and adapt our methods to keep up with the trend of technology updates.

CRediT authorship contribution statement

Junyi Li: Supervision, Project administration, Methodology, Funding acquisition, Conceptualization. **Xuan Wang:** Methodology, Conceptualization. **Yadong Wang:** Resources, Methodology. **Rui Han:**

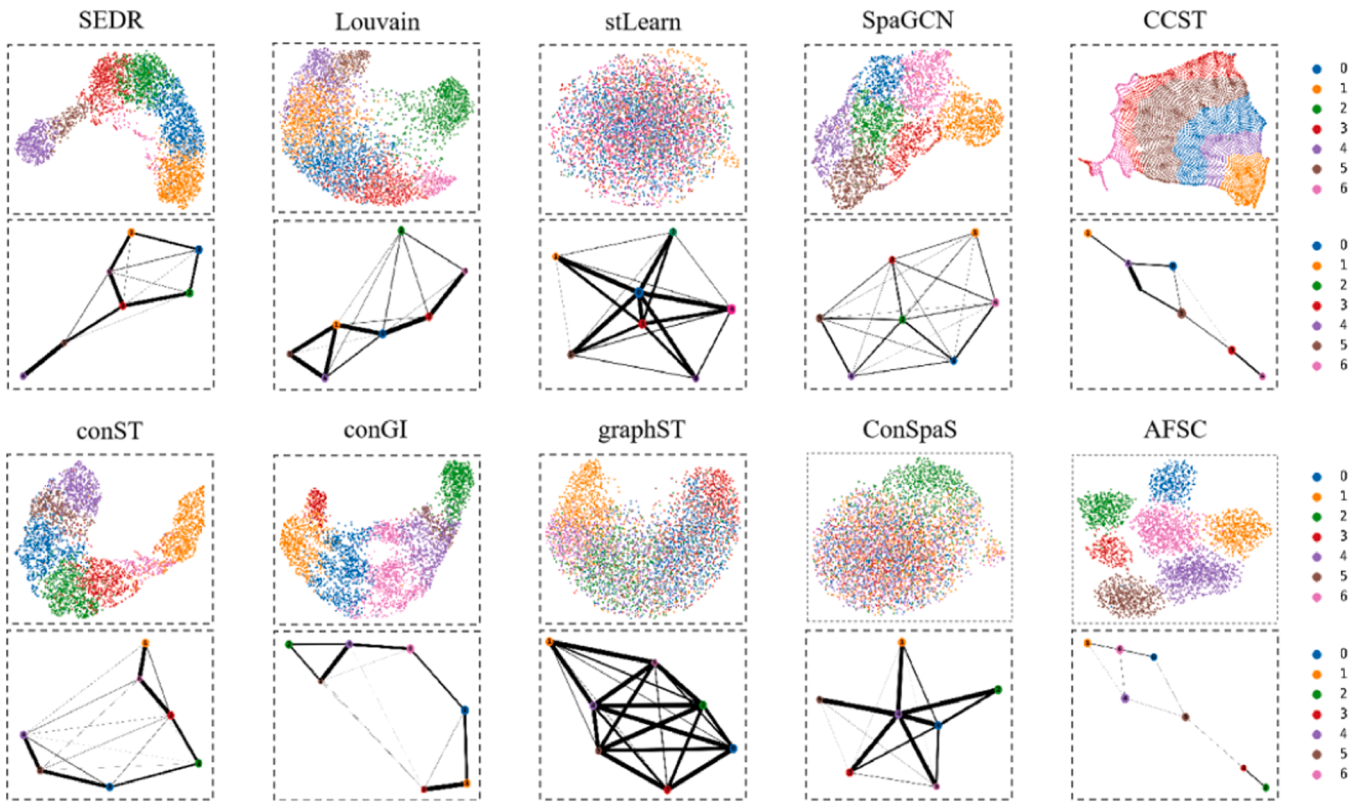

Fig. 6. UMAP visualizations and PAGA graphs. This figure shows UMAP visualizations and PAGA graphs using representations generated by methods applied to the tissue slice 151676 of DLPFC dataset. High-resolution images are available for viewing at https://github.com/bioszhr/AFSC/tree/main/results/figures.

Table 1
Ablation study results on the tissue slice 151676 of the DLPFC dataset.

Model	151676		
	ARI	NMI	FMI
Siamese($\tau = 0$)	0.429	0.584	0.521
Siamese($\tau = 1$)	0.555	0.655	0.628
w/ SGD	0.502	0.616	0.582
random_pos	0.334	0.454	0.441
knn_pos	0.561	0.604	0.602
w/o L_cluster	0.550	0.633	0.623
AFSC	**0.562**	**0.661**	**0.635**

Validation, Software, Methodology, Formal analysis, Data curation. **Xu Wang:** Validation, Software, Data curation.

Declaration of Competing Interest

The authors declare that they have no competing interests.

Acknowledgments

This work was supported by the grants from the National Key Research and Development Program of China (2021YFA0910700), National Natural Science Foundation of China (32470704), Shenzhen Science and Technology University Stable Support Program (GXWD20201230155427003-20200821222112001), Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies (2022B1212010005).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2024.09.005.

References

[1] Ji AL, Rubin AJ, Thrane K, Jiang S, Reynolds DL, Meyers RM, et al. Multimodal analysis of composition and spatial archi-tecture in human squamous cell carcinoma. Cell 2020;182(2):497–514.
[2] Lubeck E, Coskun AF, Zhiyentayev T, Ahmad M, Cai L. Single-cell in situ RNA profiling by sequential hybridization. Nat Methods 2014;11(4):360–1.
[3] Shah S, Lubeck E, Zhou W, Cai L. In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus. Neuron 2016;92 (2):342–57.
[4] Eng CHL, Lawson M, Zhu Q, Dries R, Koulena N, Takei Y, et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. Nature 2019;568(7751): 235–9.
[5] Chen KH, Boettiger AN, Moffitt JR, Wang S, Zhuang X. Spatially resolved, highly multiplexed RNA profiling in single cells. Science 2015;348(6233):aaa6090.
[6] Moffitt JR, Bambah-Mukku D, Eichhorn SW, Vaughn E, Shekhar K, Perez JD, et al. Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. Science 2018;362(6416):eaau5324.
[7] Wang X, Allen WE, Wright MA, Sylwestrak EL, Samusik N, Vesuna S, et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. Science 2018;361(6400):eaat5691.
[8] Lee JH, Daugharthy ER, Scheiman J, Kalhor R, Yang JL, Ferrante TC, et al. Highly multiplexed subcellular RNA sequenc-ing in situ. science 2014;343(6177):1360–3.
[9] Ståhl PL, Salmén F, Vickovic S, Lundmark A, Navarro JF, Magnusson J, et al. Visualization and analysis of gene expression in tis-sue sections by spatial transcriptomics. Science 2016;353(6294):78–82.
[10] Rodriques SG, Stickels RR, Goeva A, Martin CA, Murray E, Vander-burg CR, et al. Slide-seq: a scalable technology for measuring genome-wide expression at high spatial resolution. Science 2019;363(6434):1463–7.
[11] Stickels RR, Murray E, Kumar P, Li J, Marshall JL, Di Bella DJ, et al. Highly sensitive spatial transcriptomics at near-cellular reso-lution with Slide-seqV2. Nat Biotechnol 2021;39(3):313–9.
[12] Vickovic S, Eraslan G, Salmén F, Klughammer J, Stenbeck L, Schapiro D, et al. High-definition spatial transcriptomics for in situ tissue profiling. Nat Methods 2019;16(10):987–90.
[13] Moncada R, Barkley D, Wagner F, Chiodin M, Devlin JC, Baron M, et al. Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarci-nomas. Nat Biotechnol 2020;38(3): 333–42.

[14] Chen WT, Lu A, Craessaerts K, Pavie B, Frigerio CS, Corthout N, et al. Spatial transcriptomics and in situ sequencing to study Alzheimer's disease. Cell 2020;182 (4):976–91.
[15] Zeng Y, Wei Z, Zhong F, Pan Z, Lu Y, Yang Y. A parameter-free deep embedded clustering method for single-cell RNA-seq data. Brief Bioinforma 2022;23(5).
16 Zeng Y, Zhou X, Rao J, Lu Y, Yang Y. Accurately clustering single-cell RNA-seq data by capturing structural relations between cells through graph convolutional network. 2020 IEEE International Con-ference on Bioinformatics and Biomedicine (BIBM). IEEE; 2020. p. 519–22.
[17] Rao J, Zhou X, Lu Y, Zhao H, Yang Y. Imputing single-cell RNA-seq data by combining graph convolution and autoencoder neural net-works. Iscience 2021;24 (5):102393.
[18] Zeng Y, Wei Z, Yu W, Yin R, Yuan Y, Li B, et al. Spatial transcriptomics prediction from histology jointly through transformer and graph neural networks. Brief Bioinforma 2022;23(5).
[19] Fu X, Sun L, Chen JY, Dong R, Lin Y, Palmiter RD, et al. Continuous polony gels for tissue mapping with high resolution and RNA capture efficiency. 03 BioRxiv 2021: 2021. 03.
[20] Zhao E, Stone MR, Ren X, Guenthoer J, Smythe KS, Pulliam T, et al. Spatial transcriptomics at subspot resolution with BayesSpace. Nat Biotechnol 2021;39 (11):1375–84.
[21] Dries R, Zhu Q, Dong R, Eng CHL, Li H, Liu K, et al. Giotto: a toolbox for integrative analysis and visualization of spatial expression data. Genome Biol 2021;22:1–31.
[22] Xu H, Fu H, Long Y, Ang KS, Sethi R, Chong K, et al. Unsupervised spatially embedded deep representation of spatial transcriptomics. Genome Med 2024;16 (1):12.
[23] Hu J, Li X, Coleman K, Schroeder A, Ma N, Irwin DJ, et al. SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. Nat Methods 2021;18 (11):1342–51.
[24] Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. arXiv Prepr arXiv 2016;1609:02907.
[25] Dong K, Zhang S. Deciphering spatial domains from spatially re-solved transcriptomics with an adaptive graph attention auto-encoder. Nat Commun 2022;13(1):1739.
[26] Li J, Chen S, Pan X, Yuan Y, Shen HB. Cell clustering for spatial transcriptomics data with graph neural networks. Nat Comput Sci 2022;2(6):399–408.
[27] Velickovic P, Fedus W, Hamilton WL, Liò P, Bengio Y, Hjelm RD. Deep graph infomax. ICLR (Poster 2019;2(3):4.
[28] He K, Fan H, Wu Y, Xie S, Girshick R. Momentum contrast for unsupervised visual representation learning. : Proc IEEE/CVF Conf Comput Vis Pattern Recognit 2020: 9729–38.
[29] Zong Y, Yu T, Wang X, Wang Y, Hu Z, Li Y. conST: an inter-pretable multi-modal contrastive learning framework for spatial transcriptomics. 01 bioRxiv 2022:2022. 01.
[30] He K, Chen X, Xie S, Li Y, Dollár P, Girshick R. Masked autoencoders are scalable vision learners. : Proc IEEE/CVF Conf Comput Vis Pattern Recognit 2022:16000–9.
[31] Long Y, Ang KS, Li M, Chong KLK, Sethi R, Zhong C, et al. Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with GraphST. Nat Commun 2023;14(1):1155.
[32] Zeng Y, Yin R, Luo M, Chen J, Pan Z, Lu Y, et al. Identifying spatial domain by adapting transcriptomics with histology through contrastive learning. Brief Bioinforma 2023;24(2):bbad048.
[33] Chen A, Liao S, Cheng M, Ma K, Wu L, Lai Y, et al. Large field of view-spatially resolved transcriptomics at nanoscale resolution. BioRxiv 2021.
[34] Liu Y, Yang M, Deng Y, Su G, Enninful A, Guo CC, et al. High-spatial-resolution multi-omics sequencing via deterministic bar-coding in tissue. Cell 2020;183(6): 1665–81.
[35] Cho CS, Xi J, Si Y, Park SR, Hsu JE, Kim M, et al. Microscopic examination of spatial transcriptome using seq-scope. Cell 2021;184(13):3559–72.
[36] Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol 2018;19:1–5.
[37] Chen X, He K. Exploring simple siamese representation learning. : Proc IEEE/CVF Conf Comput Vis Pattern Recognit 2021:15750–8.
[38] Wang C, Pan S, Hu R, Long G, Jiang J, Zhang C. Attributed graph clustering: a deep attentional embedding approach. arXiv Prepr arXiv 2019;1906(06532).
[39] Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell 2021;184(13):3573–87.
[40] Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast un-folding of communities in large networks. J Stat Mech: -ory Exp 2008;2008(10):P10008.
[41] Chen A, Liao S, Cheng M, Ma K, Wu L, Lai Y, et al. Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. e21 Cell 2021; 185:1777–92. e21.
[42] Wu S, Qiu Y, Cheng X. ConSpaS: a contrastive learning framework for identifying spatial domains by integrating local and global similarities. Brief Bioinforma 2023; 24(6).
[43] McInnes L, Healy J, Melville J. Umap: Uniform manifold approxi-mation and projection for dimension reduction. arXiv Prepr arXiv 2018;1802(03426).
[44] GRILL JB, STRUB F, ALTCHE F, Tallec C, Richemond PH, Buchatskaya E, et al. Bootstrap your own latent: a new approach to self-supervised learning. Proc IEEE/ CVF Conf 2022;10.
[45] Rand WM. Objective criteria for the evaluation of clustering methods. J Am Stat Assoc 1971;66(336):846–50.
[46] Strehl A, Ghosh J. Cluster ensembles - a knowledge reuse framework for combining multiple partitions. J Mach Learn Res 2002:583–617.
[47] Fowlkes EB, Mallows CL. A method for comparing two hierarchical clusterings. J Am Stat Assoc 1983;78(383):553.
[48] Jaccard P. Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines. Bull Soc Vaud Sci Nat 1901;37:241–72.
[49] Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 1987;20:53–65.
[50] Davies DL, Bouldin DW. A cluster separation measure. IEEE Trans Pattern Anal Mach Intell 1979:224–7.
[51] Zhu Q, Shah S, Dries R, Cai L, Yuan GC. Identification of spatially associated subpopulations by combining scRNAseq and sequential fluorescence in situ hybridization data. Nat Biotechnol 2018;36(12):1183–90.
[52] Pham D, Tan X, Xu J, Grice LF, Lam PY, Raghubar A, et al. stLearn: integrating spatial location, tissue morphology and gene expression to find cell types, cell-cell interactions and spatial trajectories within undissociated tissues. BioRxiv 2020.
[53] Avesani S, Viesi E, Alessandrì L, Motterle G, Bonnici V, Beccuti M, et al. Stardust: improving spatial transcriptomics data analysis through space-aware modularity optimization-based clustering. GigaScience 2022;11:giac075.
[54] Alessandrì L, Cordero F, Beccuti M, Arigoni M, Olivero M, Romano G, et al. rCASC: reproducible classification analysis of single-cell sequencing data. GigaScience 2019;8(9):giz105.