Multi-modal Spatial Clustering for Spatial Transcriptomics Utilizing High-resolution Histology Images

1st Bingjun Li
School of Computing
University of Connecticut
Storrs, USA
bingjun.li@uconn.edu

2nd Mostafa Karami
School of Computing
University of Connecticut
Storrs, USA
mostafa.karami@uconn.edu

3nd Masum Shah Junayed
School of Computing
University of Connecticut
Storrs, USA
masumshah.junayed@uconn.edu

4rd Sheida Nabavi
School of Computing
University of Connecticut
Storrs, USA
sheida.nabavi@uconn.edu

Abstract—Understanding the intricate cellular environment within biological tissues is crucial for uncovering insights into complex biological functions. While single-cell RNA sequencing has significantly enhanced our understanding of cellular states, it lacks the spatial context to fully comprehend the cellular environment. Spatial transcriptomics (ST) addresses this limitation by enabling transcriptome-wide profiling while preserving spatial context. One of the principal challenges in ST data analysis is spatial clustering. Modern ST sequencing procedures typically include a high-resolution histology image, which has been shown in previous studies to be closely connected to gene expression profiles. However, current spatial clustering methods often fail to fully utilize the image information, limiting their ability to capture critical spatial and cellular interactions.

In this study, we propose the spatial transcriptomics multi-modal clustering (stMMC) model, a novel contrastive learning-based deep learning approach that integrates gene expression data with histology image features through a multi-modal parallel graph autoencoder. We tested stMMC against four state-of-the-art baseline models on two public ST datasets. The experiments demonstrated the superior performance of stMMC in terms of ARI and NMI and an ablation study validated the contributions of key components.

Index Terms—spatial transcriptomics, contrastive learning, multi-modal, graph neural network, genomic data

I. INTRODUCTION

Biological tissue samples contain highly complex cellular processes, which are shaped by cell distribution patterns, cell types, cell states, composition, and cell-cell interactions [1]. Such information is crucial for understanding tissue development, repair, and responses to external signals [1], [2]. Single-cell RNA sequencing technology has evolved dramatically in recent years to be more efficient, accessible, and accurate, which enables researchers to obtain deep insights into cellular states and led to the discovery of new cell types [3]. However, while single-cell sequencing provides valuable insights, the lack of contextual information limits the understanding of how cells cohabit, interact, and communicate within their environments [4], [5].

Spatial transcriptomics (ST) addresses this gap by enabling transcriptome-wide profiling while preserving spatial context

[6], [7], which enables researchers to move beyond the scope of cell clustering to find higher-order tissue structures. Spatial clustering on ST data has become a standard first step for any downstream analysis, such as tissue anatomy visualization, finding domain-dependent biomarkers, and mapping molecular regulatory networks [8]–[11]. As the amount of ST data rapidly expands, with technologies such as Visium [12], seq-FISH+ [13], and MERFISH [14] becoming more accessible, there is a growing need for advanced spatial clustering methods. Current ST procedures also contain a high-resolution histology image. While prior studies have proven that histology image features and gene expression are closely linked [15], [16], most current analytical methods do not fully integrate spatial information and histology images with gene expression data. This limitation obstructs the clustering models from extracting critical information from the histology images, such as cell-cell interactions and spatial changes in cell states.

Given these challenges, finding patterns in gene expression profiles with spatial and image context remains one of the significant challenges in spatial transcriptomics analysis. To address this, we propose a novel contrastive learning-based deep learning model named the spatial transcriptomics multi-modal clustering (stMMC) model, which integrates both gene expression and histology image features through a parallel graph autoencoder, leveraging contrastive learning to regulate feature extraction for each modality. Our contributions of this study are: i) we propose stMMC, a novel multi-modal contrastive learning-based clustering method using high-resolution histology images for spatial clustering in spatial transcriptomics data; ii) we demonstrate that combining gene expression data and learned histology image features improves the spatial clustering performance and paves a new way for future research; iii) we conduct experiments to show the proposed method achieves superior performance on benchmark datasets.

II. RELATED WORKS

As ST sequencing has emerged as a powerful tool for mapping gene expression profiles with spatial context, researchers have proposed several methods to tackle the spatial clustering
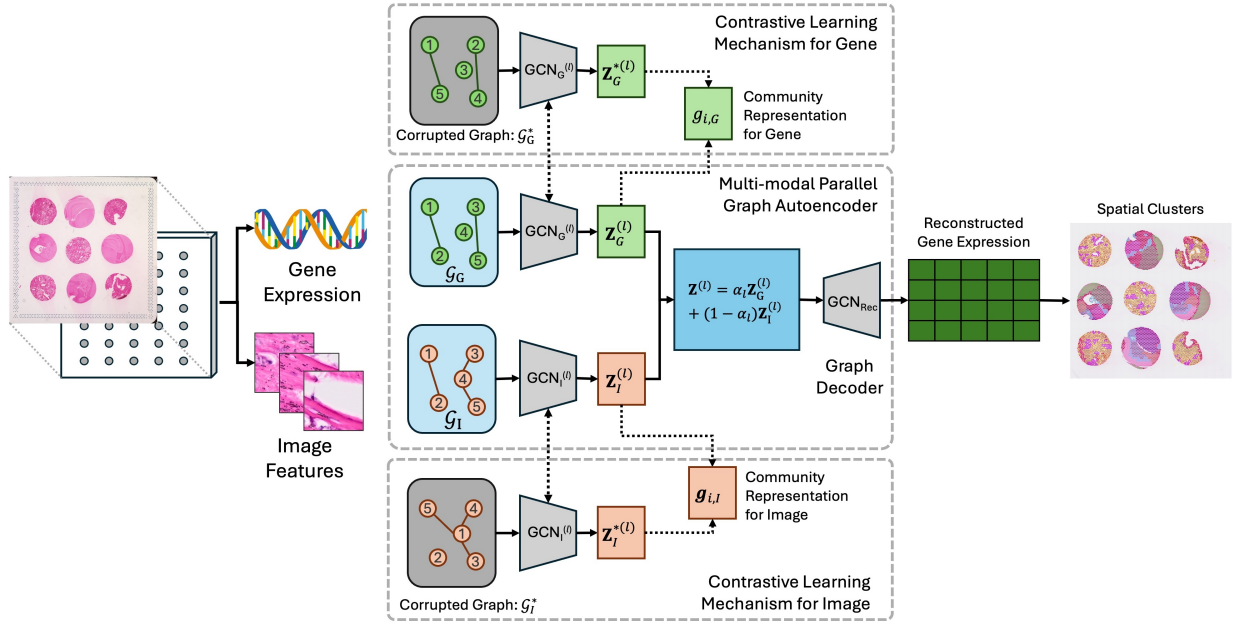
Fig. 1: The overall structure of stMMC is plotted here, where trapezoids represent the GCN layer, and rectangles represent extracted features. Two GCNs share the same weight are connected with dashed lines with double arrowheads. stMMC takes two data modalities and passes them through the multi-modal parallel graph autoencoder (MPGA), where each modality is regulated by a contrastive learning mechanism, which is shown in Figure 2. The MPGA reconstructs a refined gene expression data, which is then used for spatial clustering.

problem [17]. These methods can be divided into three major categories based on how they extract features: 1) conventional-based methods; 2) neural-network-based methods using only location information; and 3) neural-network-based methods using both location information and image features.

Zhao et al. proposed BayesSpace, which is a Bayesian statistical model that utilizes spatial neighborhood information to refine the expression data and clustering on the low-dimension representation [18]. Li et al. proposed BASS, which is able to do multi-sample and multi-scale analysis and utilizes a Bayesian hierarchical modeling framework for spatial clustering [19]. Leiden is a widely adopted general clustering model on graph-structured data [20]. It works by iteratively optimizing the partitions to identify densely connected communities within a network. These conventional-based methods are very efficient for training and inferencing, but they do not deal well with the high dimensionality of the ST data.

Most neural-network-based methods represent the location information in the form of graphs and utilize graph neural network (GNN) to extract a low-dimension representation out of the gene expression data. Ren et al. proposed SpaceFlow, which leverages GNN to integrate gene expression similarity with spatial information and clusters on the spatially consistent low-dimensional embeddings [21]. Fang et al. proposed stAA, which is an adversarial variational graph autoencoder for spatial clustering by using Wasserstein distance as the prior distribution [22]. Xu et al. proposed SEDR, which integrates learned expression features with the corresponding spatial information using a variational graph autoencoder [23].

All of these methods do not utilize the information of histology images and the few methods that do incorporate images only use tissue morphology information. Hu et al. proposed SpaGCN, which uses graph convolution networks (GCN) to combine gene expression data, location information, and an RGB value score vector from the histology images [10]. Pham et al. proposed StLearn, which utilizes spatial morphological gene expression (SME)-normalized data to perform unsupervised clustering by grouping similar points into clusters [24].

III. METHODS

As shown in Figure 1, the proposed model consists of three major modules: i) the multi-modal parallel graph autoencoder (MPGA) that consists of two independent graph autoencoders (GAEs); ii) the contrastive learning module that regulates each GAE through a contrastive learning mechanism utilizing a corrupted graph; iii) the decoder & cluster module that reconstructs the gene expression using a graph decoder and takes the refined gene expression data through a cluster to obtain the final spatial cluster assignments.

A. Problem Formulation

Consider a spatial transcriptomics dataset that has N spots with M number of gene sequencing readings, which is denoted as $\mathbf{X}_G = \left[\boldsymbol{x}_1^G, \boldsymbol{x}_2^G, ..., \boldsymbol{x}_N^G \right] \in \mathbb{R}^{N \times M}$. $M = 3000$ genes with the highest variance are selected as default for stMMC. For histology images, a square patch corresponding to each spot is extracted. An autoencoder pre-trained on ImageNet is then used to extract image features from these patches. The extracted image features are denoted as $\mathbf{X}_I \in \mathbb{R}^{N \times D}$, where D is the dimension of the image features.

To better utilize the location information along with gene expression data and histology image features, we generated a graph for each modality that incorporates spot relationship information from the other modality. For example, we created a graph for the gene expression modality using spot proximity information from the histology image modality and vice versa. Thus, we are able to fuse the information from different modalities before the fusion by aggregation. The graph for the gene expression modality is defined as $\mathcal{G}_G = (\mathbf{X}_G, V_S, E_G$, where $\mathbf{X}_G$ is the gene expression data, V_S is the set of nodes that each represents a spot on the sample, and E_G is the set of connecting edges based on the proximity between spots. The corresponding adjacency matrix is denoted as $\mathbf{A}_G$, where $\mathbf{A}_{ij} = 1$ when spot i and j are close in distance and $\mathbf{A}_{ij} = 0$ when two spots are far away. For any spot i in the $\mathcal{G}_G$, $K = 3$ nearest spots are selected to be connected. The graph for the image feature modality is defined as $\mathcal{G}_I = (\mathbf{X}_I, V_S, E_I)$, where $\mathbf{X}_I$ is the image features, V_S is the set of spots on the sample and E_I is the set of edges based on similarities between gene expression data. To compute the similarity edges, we used PCA to reduce the dimension of the gene expression data and then ran KNN based on the Euclidean distance among the gene expression features to find the $K = 3$ nearest spots for each spot. The corresponding adjacency matrix is $\mathbf{A}_I$, where $\mathbf{A}_{ij} = 1$ when the expression data from spots i and j are similar and $\mathbf{A}_{ij} = 0$ when the expression data from two spots are quite different. In summary, we create a unique graph for each modality that shares the same set of nodes, but with different node attributes and different sets of edges that carry information from the other modality.

B. Multi-modal Parallel Graph Autoencoder

To adequately extract information from each modality, two independent GAEs are used within the MPGA.

$$\mathbf{Z}_G^{(l)} = \sigma(\tilde{\mathbf{A}}_G \mathbf{Z}_G^{(l-1)} \mathbf{W}_G^{(l-1)} + \mathbf{B}_G^{(l-1)}), \qquad (1)$$

$$\mathbf{Z}_I^{(l)} = \sigma(\tilde{\mathbf{A}}_I \mathbf{Z}_I^{(l-1)} \mathbf{W}_I^{(l-1)} + \mathbf{B}_I^{(l-1)}), \qquad (2)$$

where l is the number of layers in the GAE, $\tilde{\mathbf{A}}_G$ and $\tilde{\mathbf{A}}_I$ are both the normalized adjacency matrices, $\mathbf{Z}_G^{(l)} \in \mathbb{R}^{N \times F}$ and $\mathbf{Z}_I^{(l)} \in \mathbb{R}^{N \times F}$ are the learned features from the l-th layer in the gene expression GAE and image feature GAE, respectively, F is the length of the learned features, σ is the activation function, $\mathbf{W}_G^{(l)}$ and $\mathbf{W}_E^{(l)}$ are learnable weights for the l-th layer, and $\mathbf{B}_G^{(l)}$ and $\mathbf{B}_I^{(l)}$ are the biases for the l-th layer. The features in both modality are initialized as $\mathbf{Z}_I^{(0)} = \mathbf{X}_I$ and $\mathbf{Z}_G^{(0)} = \mathbf{X}_G$. The normalized adjacency matrix for $\mathbf{A}$ is defined as $\tilde{\mathbf{A}} = \mathbf{D}^{-\frac{1}{2}} \mathbf{A} \mathbf{D}^{\frac{1}{2}}$, where $\mathbf{D}$ is the degree matrix of the graph.

The learned features from both GAEs are aggregated using weight at each layer.

$$\mathbf{Z}^{(l)} = \alpha_l \mathbf{Z}_G^{(l)} + (1 - \alpha_l) \mathbf{Z}_I^{(l)}, \qquad (3)$$

where α_l is a learnable weight for l-th layer, and L is total number of GCN layers. The aggregated feature, $Z^{(L)}$ is used as the final learned feature from MPGA and serves as the input for the graph decoder. A graph decoder is used to reconstruct the gene expression data,

$$\mathbf{X}_{Rec} = \sigma(\tilde{\mathbf{A}}_G \mathbf{Z}^{(L)} \mathbf{W}_{Rec} + \mathbf{B}_{Rec}). \qquad (4)$$

And the reconstruction loss is defined as the follow.

$$L_{Rec} = \sum_{i=1}^{N} \|\boldsymbol{x}_i^G - \boldsymbol{x}_i^{Rec}\|^2, \qquad (5)$$

where $\boldsymbol{x}_i^G \in \mathbb{R}^M$ is the gene expression data for i-th spot and $\boldsymbol{x}_i^{Rec} \in \mathbb{R}^M$ is the reconstructed expression data for i-th spot.
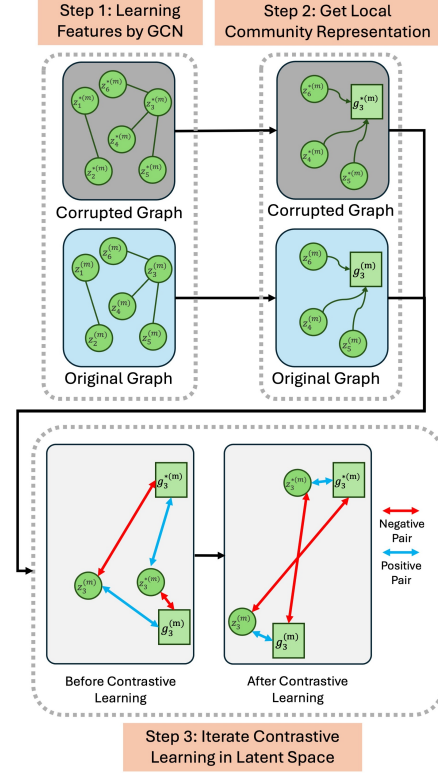

Fig. 2: The detailed process of contrastive learning mechanism for a random spot on any modality is plotted here, where the top row is the corrupted graph, the bottom row is the original graph. There are three steps of the contrastive learning mechanism: 1) obtaining the learned spot feature by GCN; 2) computing the original local community representation and the corrupted one; 3) assigning positive pairs to the learned features and the community representation from the same graph, and negative pair to the learned feature and the community representation from different graphs.

C. Contrastive Learning Module

Contrastive learning is an emerging technique used to better extract embedding features in unsupervised learning problems and has shown promising results. The core mechanism of contrastive learning is to assign positive and negative pairs to different feature embeddings, and pull the embeddings with positive pairs close and push the embeddings with negative pairs far away. Inspired by the Deep Graph Infomax approach [25], a corrupted graph is generated for each modality by shuffling nodes while maintaining the same graph topology, denoted as $\mathcal{G}_G^* = (\boldsymbol{X}_G^*, V_S^*, E_{Proximity})$ and $\mathcal{G}_I^* = (\boldsymbol{X}_I^*, V_S^*, E_{Similarity})$ for gene expression data and image

feature data, respectively. $E_{\text{Proximity}}$ and $E_{\text{Similarity}}$ stay the same during the shuffle. Corrupted graphs are fed into the same GAE within the same modality and corrupted learned features are obtained, denoted as Z_I^* and Z_G^*. To obtain the localized community information among spots, a community representation is computed for each spot with the following definition:

$$g_i^{(m)} = \frac{1}{|\text{Neb}(i)|} \sum_{j \in \text{Neb}(i)} z_{j,m}, \forall\, m \in [\text{I}, \text{G}] \qquad (6)$$

where $z_{j,m}$ is the learned representation for spot j in m-th modality, and $\text{Neb}(i)$ is the set of one-step neighbors of spot i. The learned embedding from the original graph, $z_{i,m}$ and the community representation from the original graph, $g_{i,m}$ are assigned positive pairs, while the learned embedding from the corrupted graph, $z_{i,m}^*$ and original community representation, g_i are assigned negative pairs. The key idea of the implemented contrastive learning mechanism is that the local community representation of spot i, $g_{i,m}$ should be close to the original learned embedding of the same spot, $z_{i,m}$ in the latent space, but far away from the corrupted learned embedding of the same spot, $z_{i,m}^*$. An illustration of this process for a random spot is shown in Figure 2.

A neural network-based discriminator $\Theta(\cdot)$ is used to distinguish between positive and negative pairs. $\Theta(z_{i,m}, g_{i,m})$ calculates a scalar probability score of the pair $(z_{i,m}, g_{i,m})$ being positive. The contrastive learning loss is defined based on binary cross-entropy loss as follows:

$$L_{\text{CL}} = -\frac{1}{N}\left(\sum_{m\in[\text{G},\text{I}]}\left(\sum_{i=1}^{N}\left(\mathbb{E}_{(\mathbf{X}_m,\mathbf{A}_m)}\left[\log\Theta(z_{i,m}, g_{i,m})\right]\right.\right.\right.$$
$$\left.\left.\left.+\mathbb{E}_{(\mathbf{X}_m^*,\mathbf{A}_m)}\left[\log(1-\Theta(z_{i,m}^*, g_{i,m}))\right]\right)\right)\right). \quad (7)$$

As the corrupted graph shares the same topology with the original graph, a symmetric contrastive learning loss is defined for the corrupted graph, $\mathcal{G}_m^*$ to make the overall model more stable.

$$L_{\text{CL\_C}} = -\frac{1}{N}\left(\sum_{m\in[\text{G},\text{I}]}\right.$$
$$\left(\sum_{i=1}^{N}\left(\mathbb{E}_{(\mathbf{X}_m^*,\mathbf{A}_m)}\left[\log\Theta(z_{i,m}^*, g_{i,m}^*)\right]\right.\right.$$
$$\left.\left.\left.+\mathbb{E}_{(\mathbf{X}_m^*,\mathbf{A}_m)}\left[\log(1-\Theta(z_{i,m}, g_{i,m}^*))\right]\right)\right)\right), \quad (8)$$

where $g_{i,m}^*$ is the community representation of spot i on the corrupted graph. The final total loss function is defined as follows:

$$L_{Total} = \theta_1 L_{\text{Rec}} + \theta_2 (L_{\text{CL}} + L_{\text{CL\_C}}), \qquad (9)$$

where θ_1 and θ_2 are hyperparameters for the strength of different loss terms. The default values for θ_1 and θ_2 in stMMC are 10 and 1.

D. Clustering Module

The reconstructed data is used for spatial clustering through a separate clustering module. The default clustering algorithm used in stMMC is mclust [26]. Other conventional clustering methods can also be applied. We observed that some spots

were clustered out of sync with their local spot neighborhood in stMMC's clustering assignments, which led to a deterioration in clustering performance, especially on manually annotated datasets. To address this issue, we introduced an optional smoothing step. After the clustering module produces the initial assignments, each spot is reassigned to the same cluster as the majority cluster of its nearest b neighbors. The optimal value of b is set to 50.

(a) Ground Truth

(b) stMMC

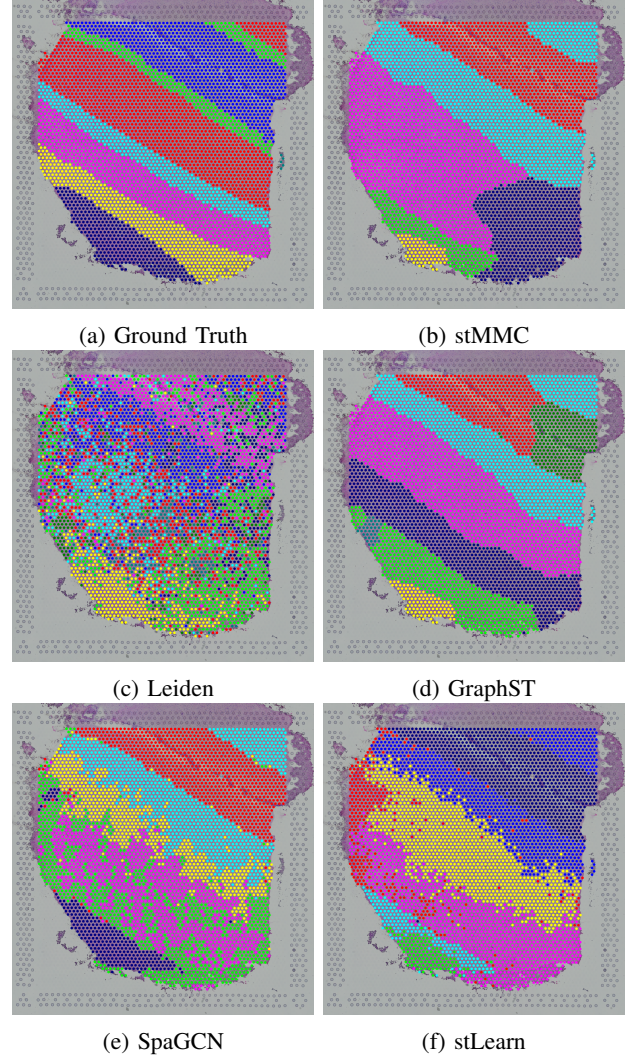(c) Leiden

(d) GraphST

(e) SpaGCN

(f) stLearn

Fig. 3: All the clustering assignments of the proposed method, stMMC and four baseline models are plotted against the ground truth for DLPFC Slice 151507.

IV. RESULTS

We tested stMMC against four state-of-the-art baseline models that cover all three of the major categories, Leiden, GraphST, stLearn, and SpaGCN [10], [20], [24], [27]. We conducted the experiments on two datasets containing 13 slice samples in total: i) DLPFC (LIBD human dorsolateral prefrontal cortex) [12], and ii) mouse_3x3_1mm from 10x Genomics [28]. The DLPFC dataset is a widely used benchmark dataset for spatial transcriptomics and is also the only one of the two datasets annotated by human experts. The dataset,

TABLE I: ARI and NMI scores of stMMC and Baseline Models on DLPFC Datasets

Slice		151507	151508	151509	151510	151669	151670	151671	151672	151673	151674	151675	151676
stMMC	ARI	**0.446**	**0.470**	**0.442**	**0.481**	**0.418**	**0.404**	**0.617**	**0.624**	**0.632**	**0.600**	**0.545**	**0.572**
	NMI	**0.652**	**0.642**	**0.646**	**0.625**	**0.560**	**0.633**	**0.711**	**0.690**	**0.727**	**0.687**	**0.652**	**0.658**
Leiden	ARI	0.150	0.146	0.114	0.098	0.116	0.065	0.153	0.086	0.205	0.210	0.239	0.226
	NMI	0.222	0.196	0.209	0.187	0.153	0.145	0.176	0.111	0.305	0.290	0.259	0.284
GraphST	ARI	0.421	0.452	0.438	0.461	0.402	0.379	0.564	0.538	0.587	0.538	0.541	0.542
	NMI	0.626	0.575	0.614	0.559	0.523	0.576	0.627	0.664	0.686	0.657	0.617	0.629
SpaGCN	ARI	0.404	0.444	0.424	0.381	0.402	0.192	0.551	0.265	0.430	0.288	0.241	0.298
	NMI	0.515	0.462	0.489	0.497	0.497	0.361	0.522	0.546	0.576	0.434	0.349	0.401
stLearn	ARI	0.425	0.294	0.414	0.266	0.337	0.188	0.279	0.337	0.311	0.315	0.446	0.359
	NMI	0.595	0.514	0.606	0.494	0.505	0.353	0.458	0.468	0.496	0.495	0.574	0.534

mouse_3x3_1mm is annotated by a graph-based clustering model in 10x Genomics Space Ranger v2.0.1. However, the histology image in mouse_3x3_1mm has a much higher resolution than that in DLPFC. While both datasets are collected using 10x genomics Visium platform with the diameter of the spot at about 55 μm, the image patches of a spot from DLPFC are about 96 pixels in width and the image patches from mouse_3x3_1mm are about 219 pixels in width. The number of clusters ranges from 6 to 9 on DLPFC slices and is 12 on the mouse_3x3_1mm dataset. stMMC is available at https://github.com/NabaviLab/stMMC. Adjusted random index (ARI) and normalized mutual information (NMI) are the metrics used in this experiment. A server with an 18-core CPU and 2 Nvidia RTX A5000 GPUs is used to run all experiments.

A. Results on Human Annotated Data

The performance of stMMC and four baseline models in terms of ARI and NMI is shown in Table I. stMMC clearly outperforms almost all four baseline models in both metrics. GraphST shows consistent performance as the second-best model across almost all slices, which is in line with the previous benchmark paper [8]. stLearn demonstrates inconsistent performance, achieving second or a very close third-best performance in about half of the sample slices, but also performing close to the worst model in some slices. The performance of SpaGCN is also inconsistent across slices but mostly worse than stLearn. Leiden performs the worst among all models as expected as it is the only model that is not designed specifically for ST data.

To better understand the clustering performance, the prediction of stMMC and baseline models for slice 151507 are mapped back onto the histology image shown in Figure 3. Compared to all baseline models, stMMC is the best at capturing major segmentation between clusters and its separation of the cluster is closer to the ground truth. GraphST captures some major segmentation but also splits larger clusters where no actual segmentation exists. Both SpaGCN and stLearn correctly capture the segmentation of major clusters in the top-right corner, but both suffer from out-of-sync small clusters.

TABLE II: ARI and NMI Scores of stMMC and Baseline Models on mouse_3x3_1mm

Model	ARI	NMI
stMMC	**0.527**	**0.761**
Leiden	0.398	0.687
GraphST	0.459	0.680
SpaGCN	0.431	0.636
stLearn	0.511	0.760

(a) Ground Truth   (b) stMMC

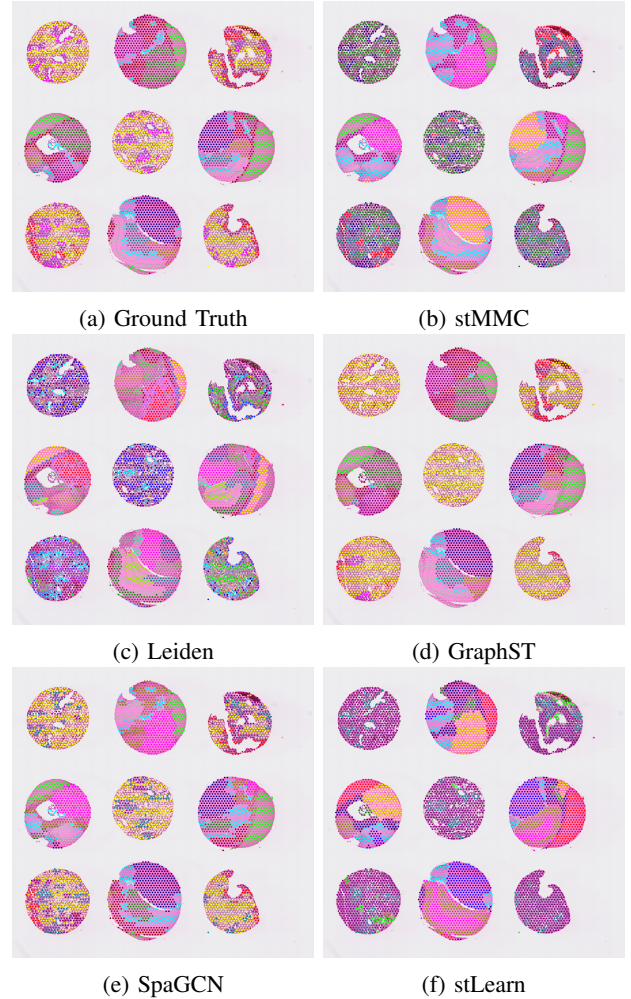(c) Leiden   (d) GraphST

(e) SpaGCN   (f) stLearn

Fig. 4: All the clustering assignments of the proposed method, stMMC and four baseline models are plotted against the ground truth for the Mouse 3x3 1mm dataset.

B. Results on Model Annotated Data

For the mouse_3x3_1mm dataset, the performance of stMMC and baseline models in terms of ARI and NMI is shown in Table II. The proposed method, stMMC clearly outperforms all four baseline models. We believe the higher resolution histology image in mouse_3x3_1mm helps stMMC achieve a better overall performance in terms of ARI and NMI. In Figure 4, the clustering assignments of stMMC and baseline models are mapped back onto the original histology image.

The segmentation by stMMC and GraphST is very close to the ground truth, while other baseline models split some of the major clusters. The segmentation lines by stMMC are closer to the grand truth compared to the ones by GraphST.

TABLE III: Ablation Study for stMMC on DLPFC 151673

Components			ARI	NMI
Smoothing	Contrastive	Image		
✓	✓	✓	0.632	0.727
✗	✓	✓	0.582	0.679
✓	✗	✓	0.586	0.668
✓	✓	✗	0.603	0.683
✗	✗	✓	0.556	0.617
✓	✗	✗	0.577	0.635
✗	✓	✗	0.571	0.631
✗	✗	✗	0.523	0.594

C. Ablation Study

To determine the contributions of each major component in stMMC, an ablation study on the contrastive learning mechanism, image feature modality, and the smoothing step is conducted on DLPFC dataset slice 151673 as shown in Table III. For variations that omit one component, stMMC without contrastive learning mechanism and stMMC without smoothing step causes the worst performance deterioration, with over 7% drop in ARI and about 8% drop in NMI, and stMMC without image feature modality causes a slightly smaller drop in performance, 4.6% drop in ARI and 6.1% drop in NMI. For variations that omit two components, stMMC without both the contrastive learning mechanism and smoothing step sees a worse performance than the other two. And stMMC without all three components is the worst performing one as expected. We can conclude that the contrastive learning mechanism and smoothing steps are the two most impactful components in stMMC and all the major components contribute as expected.

V. Conclusions

In this study, we proposed stMMC, a novel deep-learning spatial clustering method for spatial transcriptomics, which leverages a contrastive learning mechanism for better feature extraction. To the best of our knowledge, stMMC is the first method that integrates histology image features as an additional modality for spatial clustering. stMMC effectively integrates gene expression data and histology image features by using a multi-modal parallel graph autoencoder, which is proven to improve the clustering performance over state-of-the-art baseline models in experiments.

We conducted experiments with stMMC and four baseline models on two public datasets consisting of 13 sample slices in total, DLPFC and mouse_3x3_1mm. The experimental results demonstrate that stMMC consistently outperforms state-of-the-art baseline methods in terms of ARI and NMI scores. Visualization of the clustering assignments shows that stMMC is better at capturing important cluster separations. Furthermore, the ablation study validates the contributions of the three major components, the contrastive learning mechanism and the incorporation of the histology image features. These results indicate that stMMC is not only spatial clustering in spatial transcriptomics but also sets a foundation for future works that could further explore clustering in spatial transcriptomics as multi-modal data.

References

[1] M. Janiszewska, "The microcosmos of intratumor heterogeneity: the space-time of cancer evolution," *Oncogene*, 2020.

[2] M. Greaves and C. C. Maley, "Clonal evolution in cancer," *Nature*, vol. 481, no. 7381, pp. 306–313, 2012.

[3] E. Papalexi and R. Satija, "Single-cell rna sequencing to explore immune cell heterogeneity," *Nature Reviews Immunology*, 2018.

[4] A. Tanay and A. Regev, "Scaling single-cell genomics from phenomenology to mechanism," *Nature*, 2017.

[5] B. Li and S. Nabavi, "Contrastive learning in single-cell multiomics clustering," in *Proceedings of the 14th ACM BCB*, 2023.

[6] K. Vandereyken et al., "Methods and applications for single-cell and spatial multi-omics," *Nature Reviews Genetics*, Aug. 2023.

[7] M. Asp et al., "Spatially Resolved Transcriptomes—Next Generation Tools for Tissue Exploration," *BioEssays*, Oct. 2020.

[8] Z. Yuan et al., "Benchmarking spatial clustering methods with spatially resolved transcriptomics data," *Nature Methods*, Apr. 2024.

[9] Y. Chang et al., "Define and visualize pathological architectures of human tissues from spatially resolved transcriptomics using deep learning," *Computational and structural biotechnology journal*, 2022.

[10] J. Hu et al., "SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network," *Nature Methods*, Nov. 2021.

[11] D. M. Cable et al, "Cell type-specific inference of differential expression in spatial transcriptomics," *Nature methods*, 2022.

[12] K. R. Maynard et al., "Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex," *Nature neuroscience*, 2021.

[13] C.-H. L. Eng et al., "Transcriptome-scale super-resolved imaging in tissues by rna seqfish+," *Nature*, 2019.

[14] K. H. Chen et al., "Spatially resolved, highly multiplexed rna profiling in single cells," *Science*, 2015.

[15] Y. Jia et al., "THItoGene: a deep learning method for predicting spatial transcriptomics from histological images," *Briefings in Bioinformatics*.

[16] M. Pang et al., "Leveraging information in spatial transcriptomics to predict super-resolution gene expression from histology images in tumors," *BioRxiv*, 2021.

[17] M. D. Luecken et al., "Benchmarking atlas-level data integration in single-cell genomics," *Nature Methods*, Jan. 2022.

[18] E. Zhao et al., "Spatial transcriptomics at subspot resolution with BayesSpace," *Nature Biotechnology*, Nov. 2021.

[19] Z. Li and X. Zhou, "BASS: multi-scale and multi-sample analysis enables accurate cell type clustering and spatial domain detection in spatial transcriptomic studies," *Genome Biology*, Dec. 2022.

[20] V. A. Traag et al., "From louvain to leiden: guaranteeing well-connected communities," *Scientific reports*, 2019.

[21] H. Ren et al., "Identifying multicellular spatiotemporal organization of cells with SpaceFlow," *Nature Communications*, Jul. 2022.

[22] Z. Fang et al., "stAA: adversarial graph autoencoder for spatial clustering task of spatially resolved transcriptomics," *Briefings in Bioinformatics*.

[23] H. Xu et al., "Unsupervised spatially embedded deep representation of spatial transcriptomics," *Genome Medicine*, Jan. 2024.

[24] D. Pham et al., "Robust mapping of spatiotemporal trajectories and cell–cell interactions in healthy and diseased tissues," *Nature Communications*, Nov. 2023.

[25] P. Veličković et al., "Deep graph infomax," *arXiv preprint arXiv:1809.10341*, 2018.

[26] L. Scrucca et al., *Model-based clustering, classification, and density estimation using mclust in R*. Chapman and Hall/CRC, 2023.

[27] Y. Long et al., "Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with GraphST," *Nature Communications*, Mar. 2023.

[28] 10x Genomics, "Mouse tissue microarray in 3x3 layout with 1 mm edge-to-edge spacing (ffpe)," Dataset by 10x Genomics, 2024. [Online]. Available: https://www.10xgenomics.com/datasets/mouse-tissue-microarray-in-3x3-layout-with-1-mm-edge-to-edge-spacing-ffpe