SAMGTD: Spatial-Aware Masked Graph Transformer-Diffusion Model for Enhanced Cell Type Deconvolution in Spatial Transcriptomics

Shilin Zhang[1], Suixue Wang[2], Qingchen Zhang[2*], Xiulong Liu[1*]

[1]College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]School of Computer Science and Technology, Hainan University, Hainan, China
zhang_shilin_sd@163.com,{wangsuixue, zhangqingchen}@hainanu.edu.cn,xiulong_liu@tju.edu.cn

Abstract

Recent advances in spatial transcriptomics have enabled the integration of gene expression profiles with precise spatial coordinates, which have facilitated the exploration of tumor occurrence and development mechanisms, as well as the development of more effective targeted and immunotherapy approaches for tumor treatment. Deciphering cell type represents a critical challenge in spatial transcriptomics research. Existing methods are limited by the pervasive "dropout" events in spatial transcriptomics, hindering their ability to fully capture the relationship between spatial location and gene expression, thereby compromising the performance of cell type deconvolution. To address these limitations, we propose a spatial-aware masked graph transformer-diffusion model (SAMGTD) for enhanced cell type deconvolution in spatial transcriptomics. For spatial transcriptomics, the masked graph transformer model is designed to adaptively capture complex dependencies between spatial locations and gene expression. It employs a masking strategy that guides the model to focus on important local information during training, while the multi-head attention mechanism captures global context. More importantly, the spatial diffusion model is constructed to achieve the dual enhancement of spatial transcriptomics, including denoising and data imputation. It incorporates the multi-head attention mechanism and residual blocks, effectively addressing the "dropout" issue commonly encountered in spatial transcriptomics. For scRNA-seq, we construct a variational autoencoder to reduce noise interference while preserving key gene expression information. Finally, we construct a spatial-aware contrastive learning model to integrate scRNA-seq and spatial transcriptomics for cell type deconvolution. Experiments conducted on three datasets demonstrate that SAMGTD outperforms baseline methods.

Introduction

Tumors are a serious and highly prevalent disease, with malignant tumors in particular carrying an extremely high risk of mortality. Currently, tumor treatment consists mainly of chemotherapy, radiation therapy, and surgical resection. However, these treatment methods have limitations such as high side effects and unsatisfactory therapeutic results. Therefore, researchers have been committed to exploring the mechanisms of tumor occurrence and development to provide more effective targeted and immunotherapy approaches for tumor treatment.

In recent years, spatial transcriptomics has emerged that can acquire both spatial location and gene expression information (Asp, Bergenstråhle, and Lundeberg 2020)(Rao et al. 2021). This innovative technique has been used to analyze the expression patterns, interaction relationships, and signaling mechanisms of different cells within the tumor microenvironment. Through this analysis, the researchers aim to gain a deeper understanding of the characteristics and patterns of tumors.

Cell type deconvolution is a key task in spatial transcriptomics research. By integrating single-cell RNA sequencing (scRNA-seq) and spatial transcriptomics, the cell type composition of each spot can be obtained. Sun et al. proposed a spatial deconvolution algorithm (STRIDE) based on topic modeling, which used latent Dirichlet allocation to discover topics from scRNA-seq and achieve cell type estimation (Sun et al. 2022). Biancalani et al. proposed a deep learning framework (Tangram) based on nonconvex optimization for the alignment of sc/snRNA-seq and spatial transcriptomics data (Biancalani et al. 2021). However, spatial transcriptomics has a significant number of "dropout" events compared to scRNA-seq. The above methods are limited by "dropout", which has affected the performance of cell type deconvolution.

To overcome the above limitations, we propose a spatial-aware masked graph transformer-diffusion model (SAMGTD) for enhanced cell type deconvolution in spatial transcriptomics. Firstly, we implement a systematic preprocessing pipeline to ensure robust downstream analysis. Secondly, we construct the masked graph transformer model to jointly represent spatial location and gene expression information, capturing both local and global features through the masking strategy and multi-head attention. Thirdly, to overcome the "dropout" issue, the spatial diffusion model is constructed to perform denoising and data imputation for spatial transcriptomics. It incorporates residual blocks and the multi-head attention mechanism to effectively enhance the data. Fourthly, the variational autoencoder is constructed to process scRNA-seq, reducing noise and retaining key information. Finally, we construct the spatial-aware contrastive learning model to integrate scRNA-seq and spatial transcrip-

*Corresponding authors: Qingchen Zhang and Xiulong Liu.
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

tomics for cell type deconvolution. Experiments conducted on three datasets prove that SAMGTD outperforms current baseline methods.

The main contributions of this paper are as follows:

- We propose a spatial-aware masked graph transformer-diffusion model (SAMGTD) for enhanced cell type deconvolution in spatial transcriptomics, which enables integration of scRNA-seq and spatial transcriptomics while enhancing spatial transcriptomics data.
- We propose the masked graph transformer model designed to adaptively model complex spatial-gene dependencies under sparsity. It employs a masking strategy to focus on critical local features while leveraging multi-head attention to capture global features. This dual-scale context modeling is essential for robust representation learning.
- We propose the spatial diffusion model that addresses the pervasive "dropout" issue in spatial transcriptomics. The combination of multi-head attention and residual blocks within the spatial diffusion model allows for effective handling of incomplete and noisy data, ensuring better data enhancement.
- we propose the spatial-aware contrastive learning model that integrates scRNA-seq with spatial transcriptomics, and evaluate the performance of SAMGTD on three datasets.

Related Work

Graph Neural Networks

Graph neural networks (GNN), as a core methodology for processing graph-structured data, have garnered significant attention in recent years. Scarselli et al. proposed the GNN framework, which aggregates information from neighboring nodes and updates the node states to learn low-dimensional embeddings of graph-structured data (Scarselli et al. 2008). Kipf and Welling proposed the Graph Convolutional Network (GCN), designing localized filters based on spectral graph theory and directly incorporating graph structures into the inter-layer propagation process of neural networks, which enhances both model accuracy and efficiency (Kipf and Welling 2016). However, in the task of cell type deconvolution for spatial transcriptomics, graph neural networks are still in a critical exploratory stage.

Generative Model for Data Imputation

Data imputation serves as a crucial technique for addressing missing data issues, particularly in practical applications where missing data often impacts model performance and analysis results. Traditional data imputation methods (such as mean imputation, regression imputation, etc.) perform well in simple scenarios, but generative models demonstrate greater potential when dealing with complex data distributions. Generative models learn the latent data distribution to generate new sample data. Common generative models include Generative Adversarial Networks (GAN) (Yoon, Jordon, and Schaar 2018), Variational Autoencoders (VAE) (Nazabal et al. 2020), autoregressive models, and diffusion models (Tashiro et al. 2021). In the field of spatial transcriptomics, to address the "dropout" issue, we construct a spatial diffusion model for processing spatial transcriptomics data.

Cell Type Deconvolution in Spatial Transcriptomes

Cell type deconvolution is a crucial step in spatial transcriptomics data analysis. Kleshchevnikov et al. proposed a Bayesian model (cell2location) for cell type deconvolution, which employs negative binomial regression to enhance robustness in cross-dataset integration and estimates cell type references from single-cell data (Kleshchevnikov et al. 2022). Long et al. proposed a self-supervised learning-based method (GraphST) that employs graph convolutional neural networks to model gene expression patterns in spatial neighborhoods and integrates a contrastive learning framework to optimize cell-spatial mapping relationships (Long et al. 2023). Elosua-Bayes et al. proposed a cell type deconvolution method (SPOTlight) that employs seeded non-negative matrix factorization to integrate spatial transcriptomics and scRNA-seq data for inferring cell types within tissues (Shi et al. 2021). However, the above methods are limited by the "dropout" issue commonly encountered in spatial transcriptomics, which has affected the performance of cell type deconvolution.

Method

Overview of SAMGTD

Figure 1 provides an overview of the SAMGTD model. It mainly consists of six parts: data preprocessing, construction of spatial-gene graph, masked graph transformer module, spatial diffusion module, scRNA-seq denoising module, and spatial-aware contrastive learning module.

Data Preprocessing of Spatial Transcriptomics

To ensure robust downstream analysis, we implement a systematic preprocessing pipeline. Our pipeline begins with feature selection using the scanpy (V1.7.1) to identify the top 4096 highly variable genes. Normalization is subsequently performed by setting the total gene expression count of each spot to 10000. Next, the logarithmic transformation is applied to stabilize the variance in the dynamic range of gene expression values. Finally, scale to unit variance to balance the contributions of each gene during model training.

Construction of Spatial-gene Graph

To systematically model the relationship between spatial position and gene expression in spatial transcriptomics, we construct a spatial-gene graph. Specially, let the spatial-gene graph be formalized as an undirected graph $G_{spatial} = (V, E)$, where V denotes the set of nodes corresponding to spatially resolved spots; E denotes the set of edges capturing neighborhood spots relationships. In the graph $G_{spatial}$, the gene expression matrix of spots is represented as $S_{ge} = \left\{s_1, s_2, \cdots, s_{N_{sp}}\right\}$, where N_{sp} denotes the number of spots.

In the construction of the spatial-gene graph, the K-nearest neighbors (KNN) algorithm is employed to quantify the spatial proximity relationships between spots. These selected neighbors establish connectivity edges, which are

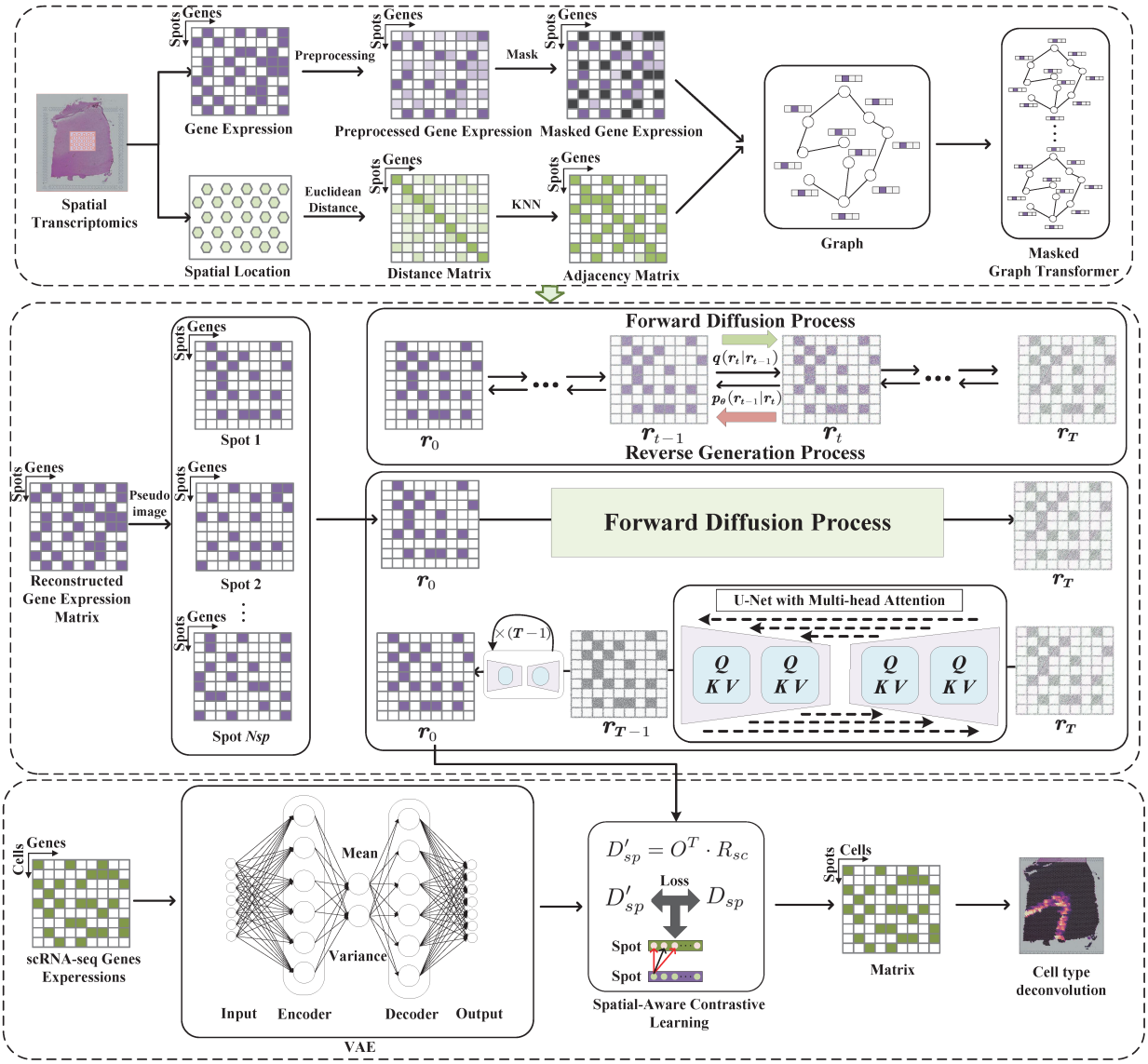

Figure 1: The overall framework of SAMGTD.

subsequently encoded into an adjacency matrix A_{sp} through binary weight assignment. A_{sp} is as shown in Formula (1).

$$a_{ij} = \begin{cases} 1, & \text{If spot } i \text{ and spot } j \text{ are close neighbors} \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where $i, j \in V$.

Masked Graph Transformer Model for Feature Encoding

To enhance the model's representation capability for spatial transcriptomics, we perform probabilistic feature masking on gene expression F_{ge}. Specifically, the binary mask matrix $Z \sim B\,(1-p)$ is generated using the Bernoulli distribution (let p = 0.4), and the gene expression matrix F_{ge} is masked by element-wise product operation:

$$S_{ge} = F_{ge} \odot Z \tag{2}$$

where $\odot$ denotes element-wise product.

Next, we construct the graph transformer model to adaptively capture the complex relationships between spatial location and gene expression (Shi et al. 2021). Specifically, given the masked gene expression matrix $S_{ge} = \left\{s_1, s_2, \cdots, s_{N_{sp}}\right\}$ of spatial transcriptomics, the attention between spot i and spot j is calculated as follows:

$$q_{h,i}^{(level)} = W_{h,q}^{(level)} s_i^{(level)} + b_{h,q}^{(level)} \tag{3}$$

$$k_{h,j}^{(level)} = W_{h,k}^{(level)} s_j^{(level)} + b_{h,k}^{(level)} \tag{4}$$

$$e_{h,ij} = W_{h,e} e_{ij} + b_{h,e} \tag{5}$$

$$\alpha_{h,ij}^{(level)} = \frac{\left\langle q_{h,i}^{(level)}, k_{h,j}^{(level)} + e_{h,ij} \right\rangle}{\sum_{y \in \mathcal{N}(i)} \left\langle q_{h,i}^{(level)}, k_{h,y}^{(level)} + e_{h,iy} \right\rangle} \tag{6}$$

where $level$ denotes the $level$-th layer, h denotes the h-th attention head, Wdenotes the weight matrix, b denotes the bias vector, $\mathcal{N}(i)$ denotes the neighbor set of spot i, and $\langle q,k\rangle = \exp\left(\frac{q^T k}{\sqrt{d}}\right)$ denotes the exponential scaled dot-product.

Then, perform aggregation using the following formula:

$$u_{h,j}^{(\text{level})} = W_{h,u}^{(\text{level})} s_j^{(\text{level})} + b_{h,u}^{(\text{level})} \tag{7}$$

$$\hat{s}_i^{(\text{level}+1)} = \|_{h=1}^{H} \left[\sum_{j \in \mathcal{N}(i)} \alpha_{h,ij}^{(\text{level})} \left(u_{h,j}^{(\text{level})} + e_{h,ij} \right) \right] \tag{8}$$

where $\|$ denotes concat operation.

To mitigate the issue of model over-smoothing, a gated residual connection is employed between network layers. The formula is as follows:

$$r_i^{(level)} = W_r^{(level)} s_i^{(level)} + b_r^{(level)} \tag{9}$$

$$\begin{aligned} \beta_i^{(level)} = \text{sigmoid}\Big(W_g^{(level)} \big[\hat{s}_i^{(level+1)}; r_i^{(level)}; \\ \hat{s}_i^{(level+1)} - r_i^{(level)} \big] \Big) \end{aligned} \tag{10}$$

$$\begin{aligned} s_i^{(level+1)} = ReLU\Big(LayerNorm\Big(\left(1-\beta_i^{(level)}\right) \\ \hat{s}_i^{(level+1)} + \beta_i^{(level)} r_i^{(level)} \Big) \Big) \end{aligned} \tag{11}$$

In the output layer, mean aggregation is used instead of the concat operation to fuse the outputs of multi-head attention (Velickovic et al. 2017), and the formula is as follows:

$$\hat{s}_i^{(level+1)} = \frac{1}{H} \sum_{h=1}^{H} \left[\sum_{j \in \mathcal{N}(i)} \alpha_{c,ij}^{(level)} \left(u_{c,j}^{(level)} + e_{c,ij}^{(level)} \right) \right] \tag{12}$$

$$s_i^{(level+1)} = \left(1-\beta_i^{(level)}\right) \hat{s}_i^{(level+1)} + \beta_i^{(level)} r_i^{(level)} \tag{13}$$

The input data is encoded through the graph transformer to generate the latent feature representation P_{sp}, which can be formalized as $E(S_{ge}, A_{sp}) \rightarrow P_{sp}$. Then, P_{sp} is reconstructed to obtain the reconstructed gene expression R_{sp}. The loss function is as follows:

$$L_{\mathrm{R}} = \sum_{i=1}^{N_{\mathrm{sp}}} \|f_i - r_i\|_F^2 \tag{14}$$

where f_i and r_i denote the original and reconstructed gene expressions, respectively.

Spatial Diffusion Model for Data Enhancement

To enhance the quality of spatial transcriptomics data and overcome the "dropout" issue, we construct the spatial diffusion model (SDM) for spatial transcriptomics. First, the 4096-dimensional gene expression vector of each spot is embedded into a 64×64 grid in a linear order, transforming it into a pseudo-image that serves as input to SDM. The core objective of SDM is to learn a model distribution $p_\theta(r_0)$ that can approximate the given true data distribution $q(r_0)$ as accurately as possible, where $r_0 \sim q(r_0)$ denotes the original data samples. The SDM is essentially a latent variable model, and its operation involves two key stochastic processes: the forward diffusion process and the reverse generation process(Ho, Jain, and Abbeel 2020)(Cui et al. 2024). The forward diffusion process gradually adds Gaussian noise to the original data r_0. As time progresses (with time steps t increasing from 1 to T), the data gradually degrades, eventually (when $t = T$) transforming into random noise r_T that follows a standard normal distribution $\mathcal{N}(0, \mathbf{I})$.

At each time step t in the forward diffusion process, noise is added to the previous state r_{t-1} to obtain the current state r_t. The conditional probability distribution is defined as:

$$q(r_t \mid r_{t-1}) = \mathcal{N}\left(r_t; \sqrt{1-\beta_t} r_{t-1}, \beta_t \mathbf{I}\right) \tag{15}$$

where $\beta_t \in (0,1)$ is a variance scheduling parameter that controls the amount of noise added at each time step, gradually corrupting the data structure. The forward diffusion process is defined as the following Markov chain:

$$q(r_{1:T} \mid r_0) = \prod_{t=1}^{T} q(r_t \mid r_{t-1}) \tag{16}$$

The important characteristic of the forward diffusion process is that r_t can be described in closed form at any time step t. Define $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, and perform the following formula derivation:

$$\begin{aligned} r_t &= \sqrt{\alpha_t} r_{t-1} + \sqrt{1-\alpha_t}\epsilon_{t-1} \\ &= \sqrt{\alpha_t}\left(\sqrt{\alpha_{t-1}} r_{t-2} + \sqrt{1-\alpha_{t-1}}\epsilon_{t-2}\right) + \sqrt{1-\alpha_t}\epsilon_{t-1} \\ &= \sqrt{\alpha_t\alpha_{t-1}} r_{t-2} + \sqrt{\sqrt{\alpha_t - \alpha_t\alpha_{t-1}}^2 + \sqrt{1-\alpha_t}^2}\,\bar{\epsilon}_{t-2} \\ &= \sqrt{\alpha_t\alpha_{t-1}} r_{t-2} + \sqrt{1-\alpha_t\alpha_{t-1}}\bar{\epsilon}_{t-2} \\ &= \sqrt{\bar{\alpha}_t} r_0 + \sqrt{1-\bar{\alpha}_t}\epsilon \end{aligned} \tag{17}$$

where $\epsilon_{t-1}, \epsilon_{t-2}, \cdots \sim \mathcal{N}(0, \mathbf{I})$.

In the framework of SDM, the reverse generation process essentially involves gradually removing noise and reconstructing the original data. If the true conditional distribution $q(r_{t-1} | r_t)$ at each time step of the reverse generative process is known exactly, it is possible to start with a random noise that follows the standard normal distribution $\mathcal{N}(0, \mathbf{I})$ and, through a series of transformations, eventually generate real samples. However, directly estimating the conditional distribution is challenging. To address this issue, we need to train a parameterized neural network model to effectively approximate it. The reverse generation process is formalized as a Markov chain(Andral et al. 2024), and the mean $\mu_\theta(r_t, t)$ and variance $\sum_\theta(r_t, t)$ of the Gaussian distribution are predicted by the neural network:

$$p_\theta(r_{0:T}) = p(r_T) \prod_{t=1}^{T} p_\theta(r_{t-1} \mid r_t) \tag{18}$$

$$p_\theta(r_{t-1} \mid r_t) = \mathcal{N}\left(r_{t-1}; \mu_\theta(r_t, t), \sum_\theta(r_t, t)\right) \tag{19}$$

where $p(r_T) = \mathcal{N}(r_T; 0, \mathbf{I})$.

SDM can be viewed as a latent variable model with T latent variables. The variational lower bound is obtained through variational inference, and maximizing it serves as the optimization objective:

$$\begin{aligned}\log p_\theta(r_0) &= \log \int p_\theta(r_{0:T})\, dr_{1:T} \\ &= \log \int \frac{p_\theta(r_{0:T})\, q(r_{1:T} \mid r_0)}{q(r_{1:T} \mid r_0)} dr_{1:T} \\ &\geq \mathbb{E}_{q(r_{1:T}|r_0)} \left[\log \frac{p_\theta(r_{0:T})}{q(r_{1:T} \mid r_0)}\right]\end{aligned} \tag{20}$$

Taking the negative of the variational lower bound, the transformation is as follows:

$$\begin{aligned}L &= -L_{\text{VLB}} \\ &= \mathbb{E}_{q(r_{1:T}|r_0)} \left[-\log \frac{p_\theta(r_{0:T})}{q(r_{1:T} \mid r_0)}\right] \\ &= \mathbb{E}_{q(r_{1:T}|r_0)} \left[\log \frac{q(r_{1:T} \mid r_0)}{p_\theta(r_{0:T})}\right]\end{aligned} \tag{21}$$

The optimization objective is further decomposed to yield the final result as follows:

$$L_{t-1}^{\text{simple}} = \mathbb{E}_{\mathbf{r}_0, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\left\| \epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}\mathbf{r}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t\right)\right\|^2\right] \tag{22}$$

In SDM, U-Net is used as the core neural network for noise prediction. During the encoder stage, the feature maps are downsampled to reduce their size and extract higher-level features. In the decoder stage, the feature maps are upsampled to restore the original resolution. More importantly, U-Net's skip connections directly link the high-resolution feature maps from the encoder layers to the corresponding layers of the decoder, preventing the loss of detailed information. The U-Net architecture incorporates residual modules and the multi-head attention mechanism, where the introduction of multi-head attention enhances global modeling capability. Through SDM processing, an enhanced spatial transcriptomics gene expression matrix D_{sp} is obtained.

VAE for scRNA-seq Denoising

For scRNA-seq, the same data preprocessing pipeline as spatial transcriptomics is adopted to obtain the preprocessed single-cell gene expression matrix F_{sc}. To reduce noise interference in scRNA-seq while preserving key gene expression information, the variational autoencoder (VAE) is constructed to process F_{sc}. Specifically, the preprocessed single-cell gene expression matrix F_{sc} is fed into the VAE. $q_\theta(z|x)$ is the encoder of VAE, where θ denotes the encoder parameters. $p_\varphi(x|z)$ is the probabilistic decoder of VAE, where φ denotes the decoder parameters. The objective function of the VAE is defined by Formula (23) (Kingma and Welling 2014).

$$\mathrm{L}_V = \mathrm{E}_{z \sim q_\theta(z|x)} \left[\log p_\varphi(x \mid z)\right] - D_{\mathrm{KL}}\left(q_\theta(z \mid x) \,|p(z)\right) \tag{23}$$

The denoised single-cell gene expression matrix R_{sc} is generated through VAE processing, which is beneficial for the subsequent cell type deconvolution.

Spatial-Aware Contrastive Learning for Cell Type Deconvolution

To achieve cell type deconvolution of spatial transcriptomics data, spatial-aware contrastive learning is constructed to integrate scRNA-seq and spatial transcriptomics. First, define a learnable deconvolution matrix $O \in \mathbb{R}^{N_{sc} \times N_{sp}}$ that represents the mapping relationship between cells and spots. The denoised single-cell gene expression matrix R_{sc} undergoes matrix operations using the mapping relationship of the deconvolution matrix to generate the predicted spatial transcriptomic gene expression matrix D'_{sp}, with calculations following Formula (24).

$$D'_{sp} = O^T \cdot R_{sc} \tag{24}$$

To learn the deconvolution matrix O, a spatial-aware contrastive learning strategy is employed for optimization. The reconstruction loss term is defined as the Mean Squared Error (MSE) computed between the predicted spatial transcriptome gene expression matrix D'_{sp} and the enhanced spatial transcriptomic gene expression matrix D_{sp}. Meanwhile, the contrastive loss term aims to maximize the similarity of neighboring spots while minimizing that of non-neighboring ones. The final objective function for learning the deconvolution matrix combines these two loss terms, as delineated in Formula (25) (Long et al. 2023).

$$\begin{aligned}L_{decon} = -\lambda_1 \sum_{i=1}^{N_{sp}} \sum_{j \in \mathcal{N}(i)} \log \frac{\exp\left(sim\left(d'_i, d_j\right)/\tau\right)}{\sum_{q \notin i}^{N_{sp}} \exp\left(sim\left(d'_i, d_q\right)/\tau\right)} \\ + \lambda_2 |D'_{sp} - D_{sp}|_F^2\end{aligned} \tag{25}$$

where $\mathcal{N}(i)$ denotes the neighbor set of spot i, λ_1 and λ_2 denotes the weight coefficient, $sim(\cdot)$ denotes the cosine similarity between two representations, τ denotes the temperature parameter.

Experiments

Datasets

In order to comprehensively evaluate the performance of SAMGTD, we conduct experiments on two real datasets and one simulated dataset (James et al. 2020)(King et al. 2021)(Park et al. 2020)(Maynard et al. 2021)(Nagy et al. 2020)(Li et al. 2022).

The Compared Methods and Evaluation Metric

To evaluate the effectiveness of SAMGTD in cell type deconvolution, we compare it experimentally with cell2location and GraphST. The area under the curve (AUC) is used as the evaluation metric and the visual comparison is performed.

Computer Platform and Implementation

The experimental environment is established on a Linux-based platform with four Intel Xeon Gold 6248R CPUs (3 GHz base frequency, 24 cores and 48 threads per CPU), and two NVIDIA A100 GPUs (80 GB memory per GPU). We extract top 4096 highly variable genes from the human

lymph node dataset and DLPFC dataset. In addition, top 256 highly variable genes are extracted from the simulation dataset as the total number of genes in this dataset is 882. The values of parameters λ_1 and λ_2 are 1 and 10 respectively. The model is employ unsupervised training.

Cell Type Deconvolution on Human Lymph Node Dataset

Figure 2: Ground Truth of the GC region in human lymph nodes.

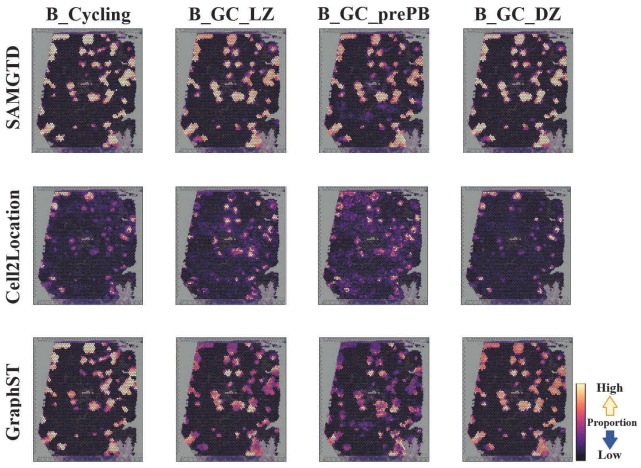

Figure 3: The visualization of human lymph nodes.

For the cell type deconvolution task within the complex microenvironment of human lymph nodes, we employ visualization and AUC evaluation metric to systematically compare the performance differences between SAMGTD and the baseline models (cell2location and GraphST). According to relevant papers published in top-tier international journals such as Nature Communications and Nature Methods (Li et al. 2022)(Li et al. 2023), cell2location, as a pivotal tool for spatial transcriptomics data analysis, maintains leading accuracy in cell type deconvolution. Notably, the recently developed GraphST algorithm by Long et al. (Long et al. 2023) integrates graph neural networks with a self-supervised learning framework, demonstrating competitive advantages over cell2location across multiple benchmark datasets. As shown in Figure 2, the study by Kleshchevnikov et al. provides ground truth for the human lymph node dataset with highlighted regions representing germinal centers (GC)(Kleshchevnikov et al. 2022). Visualizing the reconstructed spatial distribution of cell types and comparing the spatial localization of GC-associated cell types with standard GC regions can effectively reveal performance differences among methods. As shown in Figure 3, we reconstruct the spatial distribution patterns of cycling, light zone, preplasmablast, and dark zone B cells within the human lymph node GC microenvironment. The visual comparison demonstrates that SAMGTD achieves significantly better spatial consistency with ground truth in cell type localization than both cell2location and GraphST. Additionally, we evaluate the performance of different methods by calculating the AUC. As shown in Figure 4, SAMGTD achieves the highest AUC across all four cell types, demonstrating the effectiveness of our method.

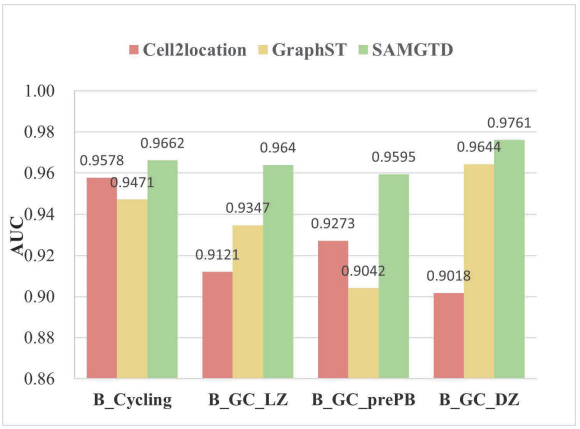

Figure 4: The AUC of human lymph nodes.

Cell Type Deconvolution on DLPFC

To systematically evaluate the performance of SAMGTD in cell type deconvolution across different datasets, we conduct experimental validation on the DLPFC dataset. As shown in Figure 5, the study by Maynard et al. revealed that DLPFC exhibits typical hierarchical characteristics (Maynard et al. 2021). We use this as the ground truth of DLPFC to evaluate the cell type deconvolution method. As shown in Figure 6, the experimental results of SAMGTD and the comparison methods are visualized separately for different cell types. Notably, SAMGTD demonstrates outstanding deconvolution capability for cortex-related cell types. Its reconstructed spatial distributions show stronger alignment with the study of Maynard, forming a clear gradient distribution from superficial to deep layers. While GraphST also captures distinct hierarchical information, cell2location's visualization results exhibit significant spatial discretization, with issues including blurred boundaries and inter-layer mixing. The experimental results of the DLPFC dataset demonstrate that SAMGTD significantly improves the effectiveness of cell type deconvolution while maintaining the continuity of tissue structure, validating the superiority in parsing complex human brain tissues.

Cell Type Deconvolution on Simulated Dataset

In addition to validating the cell type deconvolution performance on the human lymph node and DLPFC datasets, we

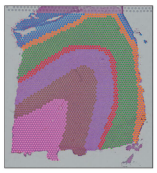

Figure 5: Ground Truth of DLPFC.

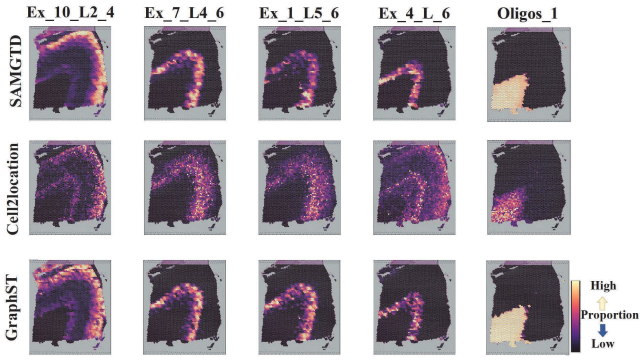

Figure 6: The visualization of DLPFC.

also conduct experimental verification on the mouse visual cortex dataset. This is a simulated dataset obtained by Li et al. through "gridding" processing of the original data (Li et al. 2022). The simulation strategy transforms the original data containing 1549 cells into 189 spots, with each spot composed of 1-18 cells while providing precise cell type information as ground truth. As shown in Figure 7, The spatial distribution of excitatory neurons generated by SAMGTD exhibits a higher degree of overlap with the ground truth compared to cell2location and GraphST.

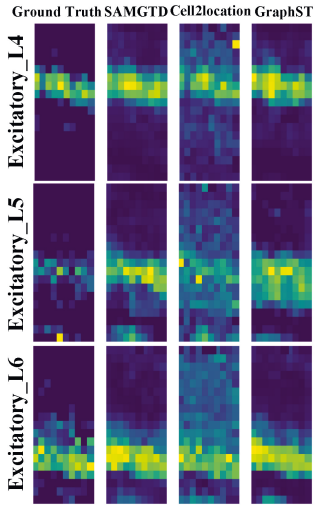

Figure 7: The visualization of simulated dataset.

Ablation Study

This ablation study aims to evaluate the contributions of key modules in the SAMGTD model to the performance of cell type deconvolution in spatial transcriptomics. Through systematic comparisons on the human lymph node dataset (evaluation metric: AUC), we validate the effectiveness of our model design. Specifically: SAMGTD (the complete model),w/o-MGT (without the Masked Graph Transformer module), w/o-SD (without the Spatial Diffusion module), w/o-VAE (without the VAE module), w/o-SCL (without the Spatial-aware Contrastive Learning module).

Method	B_Cycling	B_GC_LZ	B_GC_prePB	B_GC_DZ
w/o-SD	0.9384	0.9346	0.9116	0.9294
w/o-MGT	0.9269	0.9159	0.9183	0.9121
w/o-VAE	0.9540	0.9519	0.9462	0.9575
w/o-SCL	0.9657	0.9571	0.9528	0.9701
SAMGTD	0.9662	0.9640	0.9595	0.9761

Table 1: Ablation study.

As shown in Table 1, the experimental results demonstrate that the complete model achieves superior performance compared to all ablated versions, with the masked graph transformer module and the spatial diffusion module contributing most significantly to cell type deconvolution.

Conclusion

In this study, we propose SAMGTD, a spatial-aware masked graph transformer-diffusion model, to improve the performance of cell type deconvolution in spatial transcriptomics. One property of SAMGTD is that the masked graph transformer model adaptively captures complex dependencies between spatial locations and gene expression, enhancing feature representation through a multi-head attention mechanism and a masking strategy. More importantly, the spatial diffusion model performs dual enhancement of spatial transcriptomics. Additionally, scRNA-seq denoising and spatial-aware contrastive learning are designed to achieve cell type deconvolution. The experimental results demonstrate that SAMGTD achieves the best performance on three dataset. SAMGTD provides a tool for tumor microenvironment analysis, further facilitating the development of more effective targeted and immunotherapeutic approaches for cancer treatment.

In the future, we consider constructing a multi-modal method in combination with tissue images to extract joint representations between different modalities to further improve the performance of cell type deconvolution.

Acknowledgments

This study is supported by the grant with No. 62572157, No. 62462022 and No. KYQD(ZR)-21079.

References

Andral, C.; Douc, R.; Marival, H.; and Robert, C. P. 2024. The importance Markov chain. *Stochastic Processes and their Applications*, 171: 104316.

Asp, M.; Bergenstråhle, J.; and Lundeberg, J. 2020. Spatially resolved transcriptomes—next generation tools for tissue exploration. *BioEssays*, 42(10): 1900221.

Biancalani, T.; Scalia, G.; Buffoni, L.; Avasthi, R.; Lu, Z.; Sanger, A.; Tokcan, N.; Vanderburg, C. R.; Segerstolpe, Å.; Zhang, M.; et al. 2021. Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. *Nature Methods*, 18(11): 1352–1362.

Cui, Y.; Cui, Y.; Wang, R.; Nakai, K.; Ye, X.; Sakurai, T.; and Wei, L. 2024. DiffusionST: A diffusion model-based framework for enhancing spatial transcriptomics data quality and identifying spatial domains. *Available at SSRN 4894131*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.

James, K. R.; Gomes, T.; Elmentaite, R.; Kumar, N.; Gulliver, E. L.; King, H. W.; Stares, M. D.; Bareham, B. R.; Ferdinand, J. R.; Petrova, V. N.; et al. 2020. Distinct microbial and immune niches of the human colon. *Nature Immunology*, 21(3): 343–353.

King, H. W.; Orban, N.; Riches, J. C.; Clear, A. J.; Warnes, G.; Teichmann, S. A.; and James, L. K. 2021. Single-cell analysis of human B cell maturation predicts how antibody class switching shapes selection dynamics. *Science Immunology*, 6(56): eabe6291.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. *stat*, 1050: 1.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Kleshchevnikov, V.; Shmatko, A.; Dann, E.; Aivazidis, A.; King, H. W.; Li, T.; Elmentaite, R.; Lomakin, A.; Kedlian, V.; Gayoso, A.; et al. 2022. Cell2location maps fine-grained cell types in spatial transcriptomics. *Nature Biotechnology*, 40(5): 661–671.

Li, B.; Zhang, W.; Guo, C.; Xu, H.; Li, L.; Fang, M.; Hu, Y.; Zhang, X.; Yao, X.; Tang, M.; et al. 2022. Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution. *Nature Methods*, 19(6): 662–670.

Li, H.; Zhou, J.; Li, Z.; Chen, S.; Liao, X.; Zhang, B.; Zhang, R.; Wang, Y.; Sun, S.; and Gao, X. 2023. A comprehensive benchmarking with practical guidelines for cellular deconvolution of spatial transcriptomics. *Nature Communications*, 14(1): 1548.

Long, Y.; Ang, K. S.; Li, M.; Chong, K. L. K.; Sethi, R.; Zhong, C.; Xu, H.; Ong, Z.; Sachaphibulkij, K.; Chen, A.; et al. 2023. Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with GraphST. *Nature Communications*, 14(1): 1155.

Maynard, K. R.; Collado-Torres, L.; Weber, L. M.; Uytingco, C.; Barry, B. K.; Williams, S. R.; Catallini, J. L.; Tran, M. N.; Besich, Z.; Tippani, M.; et al. 2021. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature Neuroscience*, 24(3): 425–436.

Nagy, C.; Maitra, M.; Tanti, A.; Suderman, M.; Théroux, J.-F.; Davoli, M. A.; Perlman, K.; Yerko, V.; Wang, Y. C.; Tripathy, S. J.; et al. 2020. Single-nucleus transcriptomics of the prefrontal cortex in major depressive disorder implicates oligodendrocyte precursor cells and excitatory neurons. *Nature Neuroscience*, 23(6): 771–781.

Nazabal, A.; Olmos, P. M.; Ghahramani, Z.; and Valera, I. 2020. Handling incomplete heterogeneous data using vaes. *Pattern Recognition*, 107: 107501.

Park, J.-E.; Botting, R. A.; Domínguez Conde, C.; Popescu, D.-M.; Lavaert, M.; Kunz, D. J.; Goh, I.; Stephenson, E.; Ragazzini, R.; Tuck, E.; et al. 2020. A cell atlas of human thymic development defines T cell repertoire formation. *Science*, 367(6480): eaay3224.

Rao, A.; Barkley, D.; França, G. S.; and Yanai, I. 2021. Exploring tissue architecture using spatial transcriptomics. *Nature*, 596(7871): 211–220.

Scarselli, F.; Gori, M.; Tsoi, A. C.; Hagenbuchner, M.; and Monfardini, G. 2008. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1): 61–80.

Shi, Y.; Huang, Z.; Feng, S.; Zhong, H.; Wang, W.; and Sun, Y. 2021. Masked Label Prediction: Unified Message Passing Model for Semi-Supervised Classification. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, 1548–1554. International Joint Conferences on Artificial Intelligence Organization.

Sun, D.; Liu, Z.; Li, T.; Wu, Q.; and Wang, C. 2022. STRIDE: accurately decomposing and integrating spatial transcriptomics using single-cell RNA sequencing. *Nucleic Acids Research*, 50(7): e42–e42.

Tashiro, Y.; Song, J.; Song, Y.; and Ermon, S. 2021. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. *Advances in Neural Information Processing Systems*, 34: 24804–24816.

Velickovic, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y.; et al. 2017. Graph attention networks. *Stat*, 1050(20): 10–48550.

Yoon, J.; Jordon, J.; and Schaar, M. 2018. Gain: Missing data imputation using generative adversarial nets. In *International Conference on Machine Learning*, 5689–5698. PMLR.