Image-Enhanced Hybrid Encoding with Reinforced Contrastive Learning for Spatial Domain Identification in Spatial Transcriptomics

Daoyuan Wang[1] , **Lu Gao**[1] , **Wenlan Chen**[1] , **Cheng Liang**[2,*] , **Fei Guo**[1,*]
[1]School of Computer Science and Engineering, Central South University, Changsha 410083, China
[2]School of Information Science and Engineering, Shandong Normal University, Jinan 250358, China
wdyi701@gmail.com, 244711035@csu.edu.cn, cwlczt@163.com, alcs417@sdnu.edu.cn,
guofei@csu.edu.cn

Abstract

Spatial transcriptomics integrates spatial, gene expression, and multichannel immunohistochemistry image data, enabling advanced insights into cellular organization. However, existing methods often struggle to effectively fuse these multimodal data, limiting their potential for accurate spatial domain identification. Here, we propose IE-HERCL (Image-Enhanced Hybrid Encoding with Reinforced Contrastive Learning), a novel framework designed to address this challenge. Specifically, IE-HERCL employs hybrid encoding to capture both the non-spatial features and spatial dependencies for both gene and image modalities via autoencoders and GraphSAGE, respectively. These features are then fused using cross-view attention mechanisms to generate the unified informative embedding. To enhance the representation learning capability, we introduce a reinforced contrastive learning strategy to mitigate the influences of false negative samples, where we detect potential positive counterparts with high-order random walks. In addition, the cluster alignment is dynamically refined through optimal transport, which ensures that the fused consensus representation is coherent and robust, enabling accurate spatial domain identification. Our approach achieves state-of-the-art performance on five image-enhanced spatial transcriptomics datasets, demonstrating its robustness and effectiveness in multimodal integration and spatial domain identification. IE-HERCL offers a powerful and innovative solution for advancing spatial transcriptomics analysis. The code is released on https://github.com/wdyi701/IE-HERCL.

1 Introduction

Spatial transcriptomics has revolutionized biological research by integrating gene expression data with spatial information, offering a comprehensive view of cellular distribution and functional states within tissue microenvironments [Huang *et al.*, 2024]. Unlike traditional single-cell transcriptomics, which lacks spatial context, this approach uncovers intricate spatial patterns of gene expression [Zhu *et al.*, 2024a; Singhal *et al.*, 2024]. Its applications span tumor microenvironment analysis, tissue development, and disease pathology [Nie *et al.*, 2024; Zhong *et al.*, 2024; Cao and Gao, 2022]. Recent advancements have improved resolution from spot-level to single-cell, facilitating detailed investigations of cellular interactions and tissue organization [Zhou *et al.*, 2025; Yuan, 2024; Yang *et al.*, 2024]. However, the multimodal nature of spatial transcriptomics data—including gene expression, spatial coordinates, and tissue images—poses significant challenges for downstream analyses, particularly in spatial domain identification [Tang *et al.*, 2023a; Jia *et al.*, 2024; Jiang *et al.*, 2024]. Integrating these diverse data modalities to enhance classification performance remains a critical hurdle, limiting the broader application of this powerful tool in biological research.

Many strategies have been developed to analyze spatial transcriptomics data [Yuan *et al.*, 2024; Ma and Zhou, 2022; Long *et al.*, 2024]. Classical methods, such as Seurat and K-means, primarily cluster samples based on gene expression alone [Satija *et al.*, 2015]. While effective in identifying transcriptomic patterns, these approaches often overlook spatial relationships between cells, limiting their ability to fully interpret tissue heterogeneity. To address this limitation, advanced methods like stCluster[Wang *et al.*, 2024a], SpaCAE [Hu *et al.*, 2024] and STAGUE [Nie *et al.*, 2024] leverage graph neural networks (GNNs) to integrate gene expression and spatial information, enhancing performance by capturing spatial context . Image-based approaches, such as xSiGra, further incorporate tissue slice images to complement gene expression data, utilizing morphological and structural features [Budhkar *et al.*, 2024]. However, many of these methods rely on straightforward concatenation for multimodal integration, which may fail to account for complex interdependencies between modalities. These challenges highlight the need for a unified framework capable of effectively modeling spatial relationships, integrating multimodal data, and overcoming issues like high dimensionality and data noise.

We are motivated by three key considerations to address the challenges in spatial domain identification. First, while existing methods incorporate spatial coordinates, they often fail to fully integrate these with gene expression data. To bridge this gap, we utilize GraphSAGE [Hamilton *et al.*, 2017], a

[*]Corresponding authors

scalable and inductive graph neural network, to model spatial dependencies. By aggregating information from a fixed number of neighbors, GraphSAGE generates spatially informed cell representations, offering a robust foundation for identifying spatial domains. Second, tissue slice images contain rich morphological and structural features that complement gene expression data. Existing methods frequently rely on shallow feature extraction, overlooking deeper spatial and contextual information encoded in images. To address this, we employ a dual-encoding strategy: autoencoders capture non-spatial features, while GraphSAGE models spatial dependencies. This approach produces more discriminative and comprehensive representations of cellular environments. Third, traditional contrastive learning methods often encounter issues with false-negative samples during multimodal data fusion. To mitigate this, we introduce a reinforced contrastive learning strategy that incorporates high-order random walks to model transition probabilities between samples, reducing the impact of false negatives. Additionally, we leverage optimal transport to dynamically align clustering distributions with auxiliary distributions, further enhancing the robustness and consistency of the learned representations.

In this work, we propose Image-Enhanced Hybrid Encoding with Reinforced Contrastive Learning (IE-HERCL), a novel framework for spatial domain identification in spatial transcriptomics, as illustrated in Figure 1. IE-HERCL integrates autoencoders, graphSAGE, and attention mechanisms to extract spatial and non-spatial features from gene expression and image data. These features are fused through cross-modal attention mechanism to generate unified and consistent representation. To address challenges in multimodal data integration, we introduce a reinforced contrastive learning strategy that employs high-order random walks to mitigate the impact of false-negative samples. Additionally, optimal transport is leveraged to further enhance the robustness and reliability of the learned representations. Experimental evaluations across multiple spatial transcriptomics datasets demonstrate that IE-HERCL achieves outstanding performance in both spatial domain identification and cell type classification, consistently surpassing existing methods. This work offers an innovative and efficient framework for multimodal analysis of spatial transcriptomics data, paving the way for deeper insights into tissue organization and cellular heterogeneity.

2 Related Work

2.1 Self-Supervised Contrastive Learning

Self-supervised contrastive learning aims to bring positive pairs closer in the latent space while pushing negative pairs apart, thereby enhancing representation learning[Zeng *et al.*, 2022; Wang *et al.*, 2024b]. For example, SpaceFlow employed a spatially regularized deep graph network to encode gene expression data and spatial location information, optimizing embeddings through contrastive learning [Ren *et al.*, 2022]. Similarly, GraphST utilized graph convolutional networks (GCNs) to jointly model gene expression and spatial coordinate information [Long *et al.*, 2023]. It introduced self-supervised contrastive learning by perturbing gene expression vectors while maintaining topological structures to form negative pairs, improving latent representation learning. stDCL enhanced gene expression data to learn two latent embeddings and trained the network using spatial-aware contrastive learning and cluster-level contrastive learning to capture more discriminative representations [Yu *et al.*, 2025]. However, these methods often face challenges in effectively handling negative samples, especially when jointly modeling spatial and gene data, which can reduce model stability and accuracy. In contrast, our method utilizes high-order random walks to learn transition probabilities between samples. By weighting negative samples with very low transition probabilities, we reduce their adverse effects, enabling the model to capture meaningful patterns more effectively and achieve superior performance in spatial domain identification tasks.

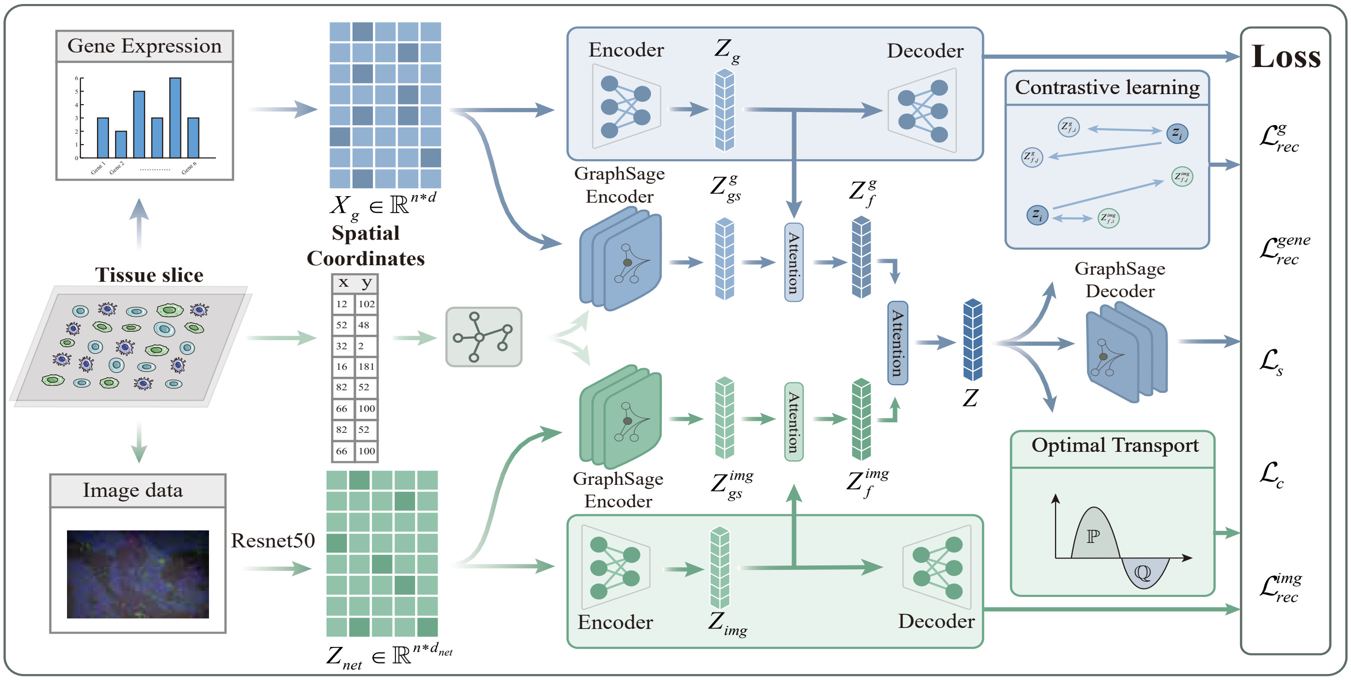
2.2 Graph Neural Networks for Spatial Domain Identification

Graph Neural Networks (GNNs) are widely applied in spatial domain identification by learning latent representations or reconstructing gene expression data [Sun *et al.*, 2025; Tang *et al.*, 2023b; Zhu *et al.*, 2024b]. For example, SpaGCN constructed adjacency graphs based on spatial coordinates and used graph convolution to aggregate information from neighboring nodes [Hu *et al.*, 2021]. The aggregated features were combined with gene expression matrices for unsupervised clustering to identify spatial domains. STAGATE employed a graph attention autoencoder to integrate spatial coordinate and gene expression data, enhancing spatial domain identification by jointly modeling spatial and transcriptomic relationships [Dong and Zhang, 2022]. STMGCN utilized a multi-graph convolutional network to encode spatial and gene expression information, employed attention mechanisms to fuse these modalities, and applied an unsupervised deep embedding clustering framework for spatial domain identification [Shi *et al.*, 2023]. MAFN adopted an end-to-end GNN to encode coordinate and gene data, combining cross-view correlation reduction strategy and attention mechanisms to learn discriminative embeddings for spatial domain identification [Zhu *et al.*, 2024c]. While these methods effectively capture relationships between spatial and transcriptomic data, they leave room for improvement in multimodal data integration, especially when combining gene expression with image data. Existing approaches often overlook deep correlations between modalities. To address this, we leverage GraphSAGE, which employ neighborhood sampling and feature aggregation strategies to generate flexible and robust node representations. Simultaneously, we incorporate image data as a complementary modality within the framework. By doing so, our method fully exploit the latent information in multimodal data, creating more discriminative embeddings for spatial domain identification.

3 Method

3.1 Preliminaries

The goal for image-enhanced spatial transcriptomics data integration is to incorporate spatial information, gene expression data, and image data to generate comprehensive latent representations. Given spatial coordinates of N points



x	y
12	102
52	48
32	2
16	181
82	52
66	100
82	52
66	100

Figure 1: An overview of the proposed IE-HERCL. The model utilizes autoencoders (AE) and GraphSAGE networks to extract features, capturing non-spatial and spatial dependencies in multimodal data. These features are fused through a triple attention mechanism to generate unified representations. The framework further optimizes the consensus representation through contrastive learning with a negative sample mitigation strategy and by refining cluster alignment using optimal transport.

$S = \{(x_i, y_i)\}_{i=1}^N$, the corresponding gene expression data $X_g \in \mathbb{R}^{N*d}$, and image data $X_{img} \in \mathbb{R}^{N*h*w*c}$, where N denotes the number of samples, d represents the dimension of gene features, and h, w, and c are the height, width, and channels of the image data, respectively. In our framework, we propose a novel approach that integrates three key innovations: first, a hybrid encoding strategy combines an autoencoder (AE) to capture non-spatial features with GraphSAGE to model spatial dependencies, effectively processing both gene and image data. Second, to mitigate the impact of false negatives in contrastive learning, we design a probability transition matrix using high-order random walks, which improves the detection of potential positive counterparts and enhances the quality of contrastive learning. Finally, we leverage optimal transport to align the auxiliary distribution of the fused consensus embedding from gene and image modalities, ensuring consistency in the representations and improving clustering accuracy. This approach enables robust multimodal data integration and significantly enhances spatial domain identification performance.

3.2 Multimodal Feature Representation Learning

In spatial transcriptomics analysis, gene expression data and image data exhibit high heterogeneity and dimensionality, making their integration challenging. To effectively extract and integrate features from these two modalities, we design a multimodal feature representation learning framework. This approach combines autoencoder (AE), GraphSAGE encoder, and cross-modal attention mechanism to extract high-quality latent representations and enhance semantic consistency between modalities.

Feature Extraction for Gene Expression Data. For gene expression data $X_g \in \mathbb{R}^{N*d}$, an AE is employed for denoising and feature learning. The AE encoder maps the input data to a latent space $Z_g \in \mathbb{R}^{N*d'}$, and the decoder reconstructs the original input. During this process, the AE removes noise, compresses redundant information, and extracts semantic features. The reconstruction loss function is defined as:

$$\mathcal{L}_{rec}^g = \left\| X_{g,i} - f_{\theta_d}^g(f_{\theta_e}^g(X_{g,i})) \right\|_F^2 \tag{1}$$

where $f_{\theta_d}^g$ and $f_{\theta_e}^g$ represent the encoder and decoder, respectively,$\|\cdot\|_F^2$ denotes the Frobenius norm to measure the difference between the input and reconstructed output. By minimizing this loss, the AE effectively extracts representative gene expression features.

Feature Extraction for Image Data. For image data $X_{img} \in \mathbb{R}^{N*h*w*c}$, high-level features $Z_{net} \in \mathbb{R}^{N*d_{net}}$ are first extracted using a pre-trained ResNet50. Subsequently, an autoencoder further processes these features to learn a deeper latent representation $Z_{img} \in \mathbb{R}^{N*d'}$, removing redundancy and capturing meaningful patterns. The reconstruction loss is defined as:

$$\mathcal{L}_{rec}^{img} = \left\| Z_{net,i} - f_{\theta_d}^{img}(f_{\theta_e}^{img}(Z_{net,i})) \right\|_F^2 \tag{2}$$

where $Z_{net,i}$ is the feature of the i-th image extracted by ResNet50, and $f_{\theta_d}^{img}$ and $f_{\theta_e}^{img}$ represent the encoder and decoder for image data, respectively. By minimizing this loss, the model extracts compact and representative image features.

Local Structure Learning. To capture local structural features in both gene expression and image data, we integrate a GraphSAGE encoder. The adjacency matrix $A \in \mathbb{R}^{N*N}$ is constructed based on spatial coordinates to define the relationships between nodes. The detailed construction process is provided in the data processing section. GraphSAGE updates node representations by aggregating information from neighboring nodes, using the following formula:

$$Z_i^v = \sigma\left(W_e \cdot \text{Concat}(X_i, \text{Aggregate}(\{X_j | j \in \mathcal{N}(i)\}))\right) \tag{3}$$

where Z_i^v represents the updated node representation, v denotes the data type (gene or image), $\mathcal{N}(i)$ is the set of neighboring nodes for node i, Aggregate$(\cdot)$ is an aggregation function (e.g., mean pooling), $\sigma(\cdot)$ is an activation function, and W_e is the learned weight matrix. GraphSAGE effectively captures local structural features and enhances spatial relationship modeling.

Cross-Modal Information Fusion. After extracting latent representations from gene expression and image data, intra-modal attention is used to adaptively fuse features within each modality. The fused representations for genes and images are given by:

$$Z_f^g = \alpha_{ae_g} Z_g + \alpha_{gs_g} Z_{gs}^g \tag{4}$$

$$Z_f^{img} = \alpha_{ae_{img}} Z_{img} + \alpha_{gs_{img}} Z_{gs}^{img} \tag{5}$$

where α_{ae} and α_{gs} are weights representing the contributions of AE and GraphSAGE encodings, respectively. Subsequently, cross-modal attention is applied to adaptively integrate the fused features from both modalities:

$$Z = \beta_g Z_f^g + \beta_{img} Z_f^{img} \tag{6}$$

where β_g and β_{img} are learnable attention weights that balance the contributions of gene and image features. This adaptive process ensures that the integrated latent representation captures consistent and robust information from both modalities.

Reconstruction of Gene Expression Data. A GraphSAGE decoder is used to reconstruct gene expression data $\hat{X}_g$ from the integrated latent representation Z. The reconstruction process is defined as:

$$\hat{X}_{g,i} = \sigma\left(W_d \cdot \text{Concat}(Z_i, \text{Aggregate}(\{Z_j | j \in \mathcal{N}(i)\}))\right) \tag{7}$$

where W_d is the learned weight matrix. The reconstruction loss is given by:

$$\mathcal{L}_{rec}^{gene} = \left\| X_{g,i} - \hat{X}_{g,i} \right\|_F^2 \tag{8}$$

The decoder reconstructs the gene expression data while preserving local structural information.

Overall Reconstruction Loss. To comprehensively consider the impact of each modality and cross-modal fusion, the total reconstruction loss is defined as:

$$\mathcal{L}_{rec} = \mathcal{L}_{rec}^g + \mathcal{L}_{rec}^{img} + \mathcal{L}_{rec}^{gene} \tag{9}$$

By minimizing the total reconstruction loss, the model learns effective multimodal latent representations, ensuring the reconstruction of accurate and consistent data.

3.3 Negative Sample Mitigation Strategy in Contrastive Learning

In contrastive learning, the selection and treatment of negative samples critically impact model performance. Conventional approaches often treat all other samples as negative for a given anchor, which can introduce false negatives (FNs) in scenarios without clear class boundaries. These FNs—samples that are incorrectly assumed to be dissimilar—can hinder the ability of the model to learn meaningful representations. To address this issue, particularly in the context of gene expression data, we propose a random walk-based negative sample mitigation strategy. This approach is designed to adjust the influence of negative samples dynamically, leveraging structural relationships among data points.

Constructing the Transition Probability Matrix via Random Walks. Given the high-dimensional and structured nature of gene expression data, we first construct an affinity graph where each sample is a node, and edge weights reflect pairwise similarities. Specifically, we use a heat kernel function to define the adjacency matrix A:

$$A_{ij} = \exp(-\|x_i - x_j\|^2 / \sigma\ ) \tag{10}$$

where x_i and x_j are the embeddings of samples i and j, and σ is a smoothing parameter that controls the decay rate of similarity. This affinity graph captures the inherent relationships among samples. To explore high-order neighborhood relationships, we perform a random walk on the graph. The random walk is represented by the transition matrix M:

$$M = D^{-1} A \tag{11}$$

where D is the degree matrix. By iterating this process, we compute a high-order transition probability matrix T:

$$T = \alpha I + (1 - \alpha) M^t \tag{12}$$

where α is a balance parameter (default 0.5) controlling the contribution of direct and high-order neighbors, and t is the number of steps in the random walk. The resulting T_{ij} encodes the probability that node j is a high-order neighbor of node i, effectively capturing distant relationships in the data. This transition matrix allows us to reduce the likelihood of incorrectly treating semantically similar samples as negatives.

Incorporating Transition Probabilities into Contrastive Learning. The transition probability matrix T is integrated into the contrastive learning framework to mitigate the impact of false negatives. In contrastive learning, views of the same sample are treated as positives, while other samples are treated as negatives. However, a naive random selection of negatives may introduce FNs in datasets without clear categorical separation. To address this, we use T_{ij} to weight negative samples during the optimization process. For a consensus representation Z_i and a view-specific representation Z_j^v, we define the contrastive loss as:

$$\mathcal{L}_c = -\frac{1}{2N} \sum_{i=1}^{N} \sum_{v=1}^{V} \log \frac{e^{\cos(Z_i, Z_i^v)/\tau}}{\sum_{j=1}^{N} \mathrm{e}^{(1-T_{ij})\cos(Z_i, Z_j^v)/\tau} - e^{1/\tau}} \tag{13}$$

where $\cos(Z_i, Z_i^v)$ denotes the cosine similarity between Z_i and Z_i^v, τ is a temperature parameter (default 0.5), and T_{ij} represents the transition probability between samples i and j. Negative samples with lower transition probabilities are assigned reduced weights, ensuring that semantically distant samples have a minimal impact on the loss. This strategy effectively mitigates the influence of false negatives and improves the robustness of the learned representations.

3.4 Optimal Transport-Based Representation Optimization

To further enhance the quality of the unified embedding, particularly in terms of achieving superior clustering performance, we propose an optimal transport (OT)-based self-optimization strategy. By leveraging a mutual supervision mechanism, this strategy dynamically aligns the clustering distribution with an auxiliary distribution, enabling adaptive clustering optimization.

Construction of Clustering and Auxiliary Distributions. We define two key distributions: the clustering distribution Q and the auxiliary distribution P. The clustering distribution Q represents the relationship between the latent embedding Z and the cluster centroids. The element q_{iu} is calculated as:

$$q_{iu} = \frac{\left(1 + \|z_i - \mu_u\|^2\right)^{-1}}{\sum_k \left(1 + \|z_i - \mu_k\|^2\right)^{-1}} \tag{14}$$

where q_{iu} indicates the probability that sample i belongs to the μ_u-th cluster. z_i is the embedding of sample i, μ_u is the u-th cluster centroid computed using pseudo-labels, and k is the total number of clusters. This distribution quantifies the probability of each sample belonging to different clusters. To enhance the contribution of high-confidence samples in clustering optimization, we construct the auxiliary distribution P as:

$$p_{iu} = \frac{q_{iu}^2 / \sum_i q_{iu}}{\sum_k \left(q_{ik}^2 / \sum_i q_{ik}\ \right)} \tag{15}$$

This design amplifies the influence of high-confidence samples, allowing them to play a more prominent role during optimization and improving the robustness of the clustering process.

Optimal Transport Objective. To align the clustering distribution Q with the auxiliary distribution P, we frame the problem as an optimal transport task. The objective is to minimize the cost of transporting mass from Q to P, while ensuring smoothness through entropy regularization. The objective function is defined as:

$$\begin{aligned} &OT(C, Q, P)^{\varepsilon} = \min_{\gamma} \langle \gamma, C \rangle_F + \varepsilon \cdot \sum_{i,j} \gamma_{i,j} \log\left(\gamma_{i,j}\right), \\ &\text{s.t.} \textstyle\sum_j \gamma_{ij} = Q_i; \sum_i \gamma_{ij} = P_j; \gamma_{ij} \geq 0, \end{aligned} \tag{16}$$

where C is the cost matrix representing the Euclidean distance between points in Q and P, $\gamma_{i,j}$ denotes the transport plan between the i-th element of Q and the j-th element of P, and ε controls the strength of entropy regularization. To measure the alignment between distributions efficiently, we employ the Sinkhorn divergence [Yu *et al.*, 2024], which is defined as:

$$\begin{aligned} \mathcal{L}_s = S_{\varepsilon}(Q, P) &:= OT(M, Q, P)^{\varepsilon} \\ &-\tfrac{1}{2}\left(OT(M, Q, Q)^{\varepsilon} + OT(M, P, P)^{\varepsilon}\right) \end{aligned} \tag{17}$$

Algorithm 1 IE-HERCL

```
Input: Multi-modal data including RNA sequencing and image data, and spatial coordinates $S$.
Parameters: Total epochs $E$, number of random walk steps $t$, and weight parameters $\lambda_1, \lambda_2, \lambda_3$.
Initialize the spatial coordinate graph $G$ using distance nearest neighbors.
for epoch in $1, 2, \ldots, E$ do
    Calculate $\mathcal{L}_{rec}^{g}$ for the gene modality using Eq. (1).
    Calculate $\mathcal{L}_{rec}^{img}$ for the image modality using Eq. (2).
    Obtain modality-specific encodings $Z_f^g$ and $Z_f^{img}$.
    Map all modality-specific encodes to a unified embedding $Z$ using Eq. (6).
    Update AE and GraphSAGE encoders and decoders using Eq. (9).
    Calculate the contrastive learning loss with a negative sample mitigation strategy using Eq. (13).
    Calculate the optimal transport loss $\mathcal{L}_s$ using Eq. (17).
    Train IE-HERCL by minimizing $\mathcal{L}$ using Eq. (18).
end for
Output: Unified embedding $Z$ and reconstructed gene expression data $\hat{X}$.
```

In practice, the auxiliary distribution P is constructed based on the clustering distribution Q. By continuously optimizing the alignment between Q and P, high-confidence data points are expected to play a dominant role in the clustering process. This optimization continues iteratively until a predefined maximum number of iterations is reached, achieving deep refinement of the consistency representation and enhancing the performance of the model in clustering assignment tasks.

3.5 Overall Optimization Objective

The final loss function of IE-HERCL integrates three components: the reconstruction loss $\mathcal{L}_{rec}$, the contrastive learning loss $\mathcal{L}_c$, and the optimal transport loss $\mathcal{L}_s$. The joint optimization objective is given as:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{rec} + \lambda_2 \mathcal{L}_c + \lambda_3 \mathcal{L}_s \tag{18}$$

where λ_1, λ_2, and λ_3 are the weight parameters for each loss term. By jointly optimizing these objectives, the proposed method effectively integrates spatial information, image data, and gene expression data to learn a more consensus embedding representation, significantly improving clustering performance. Algorithm 1 provides a detailed description of the entire model workflow.

4 Experiments

4.1 Experimental Setups

Datasets. We evaluate the proposed method on five widely used spatial transcriptomics datasets that include image information: (1) NanoString Lung 9-1, (2) 10x Visium DLPFC,

	NanoString Lung 9-1					Human breast cancer				
Methods	ARI	NMI	AMI	FMI	HS	ARI	NMI	AMI	FMI	HS
scanpy	0.2918	0.4554	0.4553	0.4443	0.4192	0.5343	0.6395	0.6327	0.5696	0.6394
STAGATE	0.3658	0.4087	0.4086	0.5095	0.3822	0.4448	0.6613	0.6552	0.4866	0.6379
GraphST	0.4084	0.4421	0.442	0.5493	0.4237	0.3571	0.6881	0.6428	0.3870	0.6659
SiGra	0.5216	0.5313	0.5313	0.6353	0.5100	0.5030	0.6141	0.6070	0.5395	0.6092
xSiGra	0.4377	0.4360	0.4359	0.5690	0.4162	0.6026	0.6782	0.6723	0.6322	0.6652
stDCL	0.3110	0.4470	0.4469	0.4736	0.4287	0.5813	0.6889	0.6839	0.6145	0.7059
IE-HERCL	**0.5679**	**0.5428**	**0.5427**	**0.6529**	**0.5348**	**0.6279**	**0.7133**	**0.7081**	**0.6757**	**0.7207**

	Mouse anterior brain					Mouse coronal brain				
Methods	ARI	NMI	AMI	FMI	HS	ARI	NMI	AMI	FMI	HS
scanpy	0.3428	0.6589	0.6102	0.3697	0.6428	0.4952	0.6084	0.6070	0.5770	0.6189
STAGATE	0.3939	0.7274	0.6878	0.4242	0.7032	0.4231	0.5810	0.5795	0.5145	0.5869
GraphST	0.3840	0.6884	0.6438	0.4097	0.6714	0.4288	0.5386	0.5370	0.5246	0.5514
SiGra	0.4687	0.6729	0.6266	0.4894	0.6626	0.3278	0.5085	0.5067	0.4519	0.5358
xSiGra	0.3440	0.6739	0.6274	0.3715	0.6570	0.5374	0.6309	0.6196	0.6127	0.6346
stDCL	0.4686	0.7212	0.6892	0.4897	0.7344	0.5286	0.5874	0.5859	0.6115	0.6150
IE-HERCL	**0.5255**	**0.7277**	**0.6976**	**0.5500**	**0.7559**	**0.5405**	**0.6373**	**0.6361**	**0.6225**	**0.6560**

Table 1: Clustering performance of all methods on NanoString Lung 9-1, Human breast cancer, Mouse brain anterior and Mouse brain coronal datasets.

(3) 10x Visium human breast cancer data, (4) mouse brain anterior slice, and (5) mouse brain coronal slice. Details on the dataset processing can be found in the Supplementary Material.

Baseline Methods. To validate the superiority of our model, we compare it with recent state-of-the-art methods, including four single-modal spatial domain identification approaches, namely Scanpy [Wolf *et al.*, 2018], STAGATE [Dong and Zhang, 2022], GraphST [Long *et al.*, 2023] and stDCL [Yu *et al.*, 2025], as well as two image-enhanced spatial domain identification methods, SiGra [Tang *et al.*, 2023a] and xSiGra [Budhkar *et al.*, 2024]. A detailed description of these methods can be found in the supplementary materials.

Evaluation Metrics. To comprehensively evaluate the performance of spatial domain identification, we adopt five widely used clustering metrics: Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), Adjusted Mutual Information (AMI), Fowlkes-Mallows Index (FMI), and Homogeneity Score (HS) [Long *et al.*, 2024]. These metrics collectively capture clustering quality from multiple perspectives, ensuring a thorough assessment of model performance.

Implementation Details. The entire model is implemented using PyTorch 1.13.0 and all experiments are conducted on an Ubuntu 20.04 server with an NVIDIA 3090 GPU. For the autoencoder in our model, we employ two layers (512-dimensional to 64-dimensional), while the GraphSAGE layer is set to a single layer (64-dimensional). To optimize the model, we utilize the Adam optimizer with a weight decay of 0.0001 and an initial learning rate of 0.001. For the other comparison methods, we use the default parameters from the original papers.

4.2 Experimental Results

Results Comparison. To demonstrate the superiority of our model, we conduct both qualitative and quantitative evaluations on the datasets mentioned earlier. For the qualitative evaluation of IE-HERCL, we visualize the spatial domains of a lung cancer tissue FOV. As shown in Figure 2, our method exhibits a strong alignment with the true labels. In contrast, the comparison method, Scanpy, which does not incorporate spatial coordinate information or pathological image data, shows poor alignment between the spatial distribution and the true labels. The superior performance of IE-HERCL is primarily attributed to the incorporation of pathological image data, which enhances representation learning and enables the model to generate robust representations. In terms of quantitative metrics (Supplementary Figure 1), IE-HERCL achieves superior performance across ARI, NMI, and HS on all lung cancer and DLPFC tissue slices, with particularly notable improvements in lung cancer tissue where pathological image data is utilized. Furthermore, on other datasets, including human breast cancer, mouse brain anterior slices, and mouse brain coronal slices (as shown in Table 1), IE-HERCL outperforms all competing methods across all metrics. We also present spatial domain distributions, UMAP visualizations, and trajectory inferences for the DLPFC dataset; detailed results are provided in the supplementary materials. Notably, on the human breast cancer dataset, IE-HERCL surpasses the second-best method, stDCL, by 4.66% in ARI and 6.12% in FMI. This outstanding performance can be attributed to the reinforced contrastive learning strategy, which mitigates the influence of negative samples, and the optimal transport mechanism, which optimizes the learning of latent representations, resulting in highly competitive outcomes. For additional experimental results, please refer to the supplementary

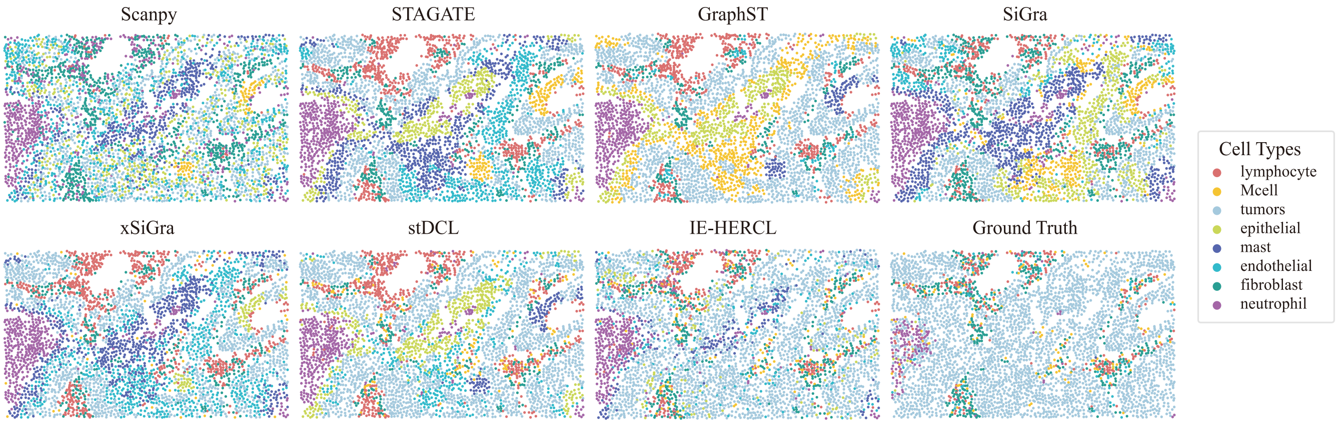


Figure 2: Spatial distribution obtained by all methods on FOV2 of the NanoString Lung 9-1 dataset.

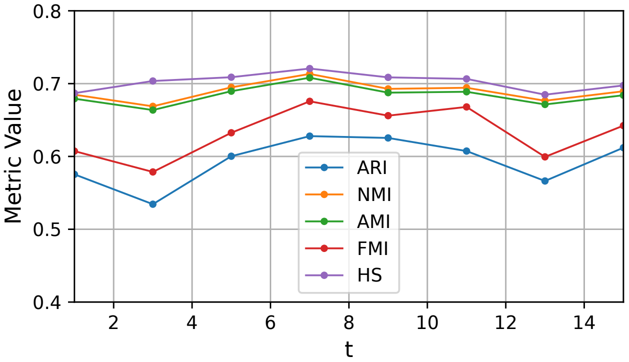


Figure 3: Performance of different random walk steps on the human breast cancer dataset.

materials.

Parameter Analysis. To further validate the sensitivity of our model to parameter settings, we conduct a series of parameter sensitivity experiments. These experiments evaluate the effects of the weight factors λ_1, λ_2 and λ_3, which are set to [1, 10], [0.0001, 0.001] and [0.1, 1], respectively, and the random walk step size t, which varies from 1 to 15 with an interval of 2. Fixing λ_1, λ_2 and λ_3 at 10, 0.001 and 1, Figure 3 demonstrates that the optimal step size t for the random walk is 9 for the above dataset. Furthermore, fixing the random walk step size at t=5, Supplementary Table 1 shows that the model achieves the best performance on the human breast cancer datasets when λ_1, λ_2 and λ_3 are set to 10, 0.001 and 1, respectively. These results suggest that high-order random walks effectively mitigate the influence of negative samples, thereby improving clustering performance.

4.3 Ablation Studies

In this section, we evaluate the contribution of each component in IE-HERCL using the human breast cancer dataset. To verify the effectiveness of contrastive learning, we remove the $\mathcal{L}_c$ module (denoted as ”w/o $\mathcal{L}_c$”). To assess the impact of optimal transport in aligning latent representations, we exclude the $\mathcal{L}_s$ module (denoted as ”w/o $\mathcal{L}_s$”). We also evaluate the effect of mitigating negative samples in contrastive learning by removing the high-order random walks (denoted as ”w/o Random Walk”). Additionally, we analyze the importance of the image data modality by removing it from the model and using only gene expression data for representation learning and clustering (denoted as ”w/o Image Data”). As shown in Table 2, each component contributes significantly to the model’s performance. The removal of any single component results in a noticeable decline in clustering performance. The best results are achieved when all components are included and the image data modality is utilized. These findings highlight the robustness and effectiveness of the proposed framework, particularly the synergy between its components.

Methods	ARI	NMI	AMI	FMI	HS
IE-HERCL	**0.6279**	**0.7133**	**0.7081**	**0.6757**	**0.7207**
w/o $\mathcal{L}_c$	0.5344	0.6688	0.6637	0.5787	0.7036
w/o $\mathcal{L}_s$	0.6120	0.6895	0.6840	0.6423	0.6975
w/o Random Walk	0.5582	0.6786	0.6733	0.5913	0.6805
w/o Image Data	0.6074	0.6943	0.6888	0.6680	0.7065

Table 2: Ablation Study on the Human Breast Cancer Dataset. w/o denotes the abbreviation without.

5 Conclusion

In this work, we present IE-HERCL, a novel framework for spatial domain identification in spatial transcriptomics that seamlessly integrates gene expression, spatial information, and histological image data. By employing a hybrid encoding strategy with autoencoders and GraphSAGE, IE-HERCL captures non-spatial features and spatial dependencies within and across multimodal data. Cross-modal attention mechanisms further fuse these features into unified representations, enabling a comprehensive understanding of tissue organization. Additionally, we introduce a reinforced contrastive learning strategy that combines high-order random walks and optimal transport to reduce the impact of false negatives on representation learning and improve clustering alignment. Extensive experiments on five diverse spatial transcriptomics datasets demonstrate the superior performance of IE-HERCL in spatial domain identification and cell type classification, consistently surpassing state-of-the-art methods.

Acknowledgments

This study was partially supported by the National Natural Science Foundation of China under Grant Nos. 62372279, 62322215, 62172296, and the Natural Science Foundation of Shandong Province under Grant No. ZR2023MF119. It was also partially supported by the high-performance computing server of Central South University.

References

[Budhkar *et al.*, 2024] Aishwarya Budhkar, Ziyang Tang, Xiang Liu, Xuhong Zhang, Jing Su, and Qianqian Song. xsigra: explainable model for single-cell spatial data elucidation. *Briefings in Bioinformatics*, 25(5):bbae388, 2024.

[Cao and Gao, 2022] Zhi-Jie Cao and Ge Gao. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nature Biotechnology*, 40(10):1458–1466, 2022.

[Dong and Zhang, 2022] Kangning Dong and Shihua Zhang. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nature communications*, 13(1):1739, 2022.

[Hamilton *et al.*, 2017] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

[Hu *et al.*, 2021] Jian Hu, Xiangjie Li, Kyle Coleman, Amelia Schroeder, Nan Ma, David J Irwin, Edward B Lee, Russell T Shinohara, and Mingyao Li. Spagcn: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nature methods*, 18(11):1342–1351, 2021.

[Hu *et al.*, 2024] Yaofeng Hu, Kai Xiao, Hengyu Yang, Xiaoping Liu, Chuanchao Zhang, and Qianqian Shi. Spatially contrastive variational autoencoder for deciphering tissue heterogeneity from spatially resolved transcriptomics. *Briefings in Bioinformatics*, 25(2):bbae016, 2024.

[Huang *et al.*, 2024] Jinjin Huang, Xiaoqian Fu, Zhuangli Zhang, Yinfeng Xie, Shangkun Liu, Yarong Wang, Zhihong Zhao, and Youmei Peng. A graph self-supervised residual learning framework for domain identification and data integration of spatial transcriptomics. *Communications Biology*, 7(1):1123, 2024.

[Jia *et al.*, 2024] Yuran Jia, Junliang Liu, Li Chen, Tianyi Zhao, and Yadong Wang. Thitogene: a deep learning method for predicting spatial transcriptomics from histological images. *Briefings in Bioinformatics*, 25(1):bbad464, 2024.

[Jiang *et al.*, 2024] Xi Jiang, Shidan Wang, Lei Guo, Bencong Zhu, Zhuoyu Wen, Liwei Jia, Lin Xu, Guanghua Xiao, and Qiwei Li. iimpact: integrating image and molecular profiles for spatial transcriptomics analysis. *Genome Biology*, 25(1):147, 2024.

[Long *et al.*, 2023] Yahui Long, Kok Siong Ang, Mengwei Li, Kian Long Kelvin Chong, Raman Sethi, Chengwei Zhong, Hang Xu, Zhiwei Ong, Karishma Sachaphibulkij, and Ao Chen. Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with graphst. *Nature Communications*, 14(1):1155, 2023.

[Long *et al.*, 2024] Yahui Long, Kok Siong Ang, Raman Sethi, Sha Liao, Yang Heng, Lynn van Olst, Shuchen Ye, Chengwei Zhong, Hang Xu, and Di Zhang. Deciphering spatial domains from spatial multi-omics with spatialglue. *Nature Methods*, pages 1–10, 2024.

[Ma and Zhou, 2022] Ying Ma and Xiang Zhou. Spatially informed cell-type deconvolution for spatial transcriptomics. *Nature biotechnology*, 40(9):1349–1359, 2022.

[Nie *et al.*, 2024] Wan Nie, Yingying Yu, Xueying Wang, Ruohan Wang, and Shuai Cheng Li. Spatially informed graph structure learning extracts insights from spatial transcriptomics. *Advanced Science*, 11(45):2403572, 2024.

[Ren *et al.*, 2022] Honglei Ren, Benjamin L Walker, Zixuan Cang, and Qing Nie. Identifying multicellular spatiotemporal organization of cells with spaceflow. *Nature communications*, 13(1):4076, 2022.

[Satija *et al.*, 2015] Rahul Satija, Jeffrey A Farrell, David Gennert, Alexander F Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. *Nature biotechnology*, 33(5):495–502, 2015.

[Shi *et al.*, 2023] Xuejing Shi, Juntong Zhu, Yahui Long, and Cheng Liang. Identifying spatial domains of spatially resolved transcriptomics via multi-view graph convolutional networks. *Briefings in Bioinformatics*, 24(5):bbad278, 2023.

[Singhal *et al.*, 2024] Vipul Singhal, Nigel Chou, Joseph Lee, Yifei Yue, Jinyue Liu, Wan Kee Chock, Li Lin, Yun-Ching Chang, Erica Mei Ling Teo, and Jonathan Aow. Banksy unifies cell typing and tissue domain segmentation for scalable spatial omics data analysis. *Nature Genetics*, 56(3):431–441, 2024.

[Sun *et al.*, 2025] Xue Sun, Wei Zhang, Wenrui Li, Na Yu, Daoliang Zhang, Qi Zou, Qiongye Dong, Xianglin Zhang, Zhiping Liu, and Zhiyuan Yuan. Spagra: Graph augmentation facilitates domain identification for spatially resolved transcriptomics. *Journal of Genetics Genomics*, 52(1):93–104, 2025.

[Tang *et al.*, 2023a] Ziyang Tang, Zuotian Li, Tieying Hou, Tonglin Zhang, Baijian Yang, Jing Su, and Qianqian Song. Sigra: single-cell spatial elucidation through an image-augmented graph transformer. *Nature Communications*, 14(1):5618, 2023.

[Tang *et al.*, 2023b] Ziyang Tang, Tonglin Zhang, Baijian Yang, Jing Su, and Qianqian Song. spaci: deciphering spatial cellular communications through adaptive graph model. *Briefings in Bioinformatics*, 24(1):bbac563, 2023.

[Wang *et al.*, 2024a] Tao Wang, Han Shu, Jialu Hu, Yongtian Wang, Jing Chen, Jiajie Peng, and Xuequn Shang. Accurately deciphering spatial domains for spatially resolved transcriptomics with stcluster. *Briefings in Bioinformatics*, 25(4):bbae329, 2024.

[Wang *et al.*, 2024b] Tianqi Wang, Huitong Zhu, Yunlan Zhou, Weihong Ding, Weichao Ding, Liangxiu Han, and Xueqin Zhang. Graph attention automatic encoder based on contrastive learning for domain recognition of spatial transcriptomics. *Communications Biology*, 7(1):1351, 2024.

[Wolf *et al.*, 2018] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19:1–5, 2018.

[Yang *et al.*, 2024] Wenyi Yang, Pingping Wang, Shouping Xu, Tao Wang, Meng Luo, Yideng Cai, Chang Xu, Guangfu Xue, Jinhao Que, and Qian Ding. Deciphering cell–cell communication at single-cell resolution for spatial transcriptomics with subgraph-based graph attention network. *Nature Communications*, 15(1):7101, 2024.

[Yu *et al.*, 2024] Jixiang Yu, Nanjun Chen, Ming Gao, Xiangtao Li, and Ka-Chun Wong. Unsupervised gene-cell collective representation learning with optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 356–364, 2024.

[Yu *et al.*, 2025] Zhuohan Yu, Yuning Yang, Xingjian Chen, Ka-Chun Wong, Zhaolei Zhang, Yuming Zhao, and Xiangtao Li. Accurate spatial heterogeneity dissection and gene regulation interpretation for spatial transcriptomics using dual graph contrastive learning. *Advanced Science*, 12(3):e2410081, 2025.

[Yuan *et al.*, 2024] Zhiyuan Yuan, Fangyuan Zhao, Senlin Lin, Yu Zhao, Jianhua Yao, Yan Cui, Xiao-Yong Zhang, and Yi Zhao. Benchmarking spatial clustering methods with spatially resolved transcriptomics data. *Nature Methods*, 21(4):712–722, 2024.

[Yuan, 2024] Zhiyuan Yuan. Mender: fast and scalable tissue structure identification in spatial omics data. *Nature Communications*, 15(1):207, 2024.

[Zeng *et al.*, 2022] Yuansong Zeng, Rui Yin, Mai Luo, Jianing Chen, Zixiang Pan, Yutong Lu, Weijiang Yu, and Yuedong Yang. Deciphering spatial domains by integrating histopathological image and tran-scriptomics via contrastive learning. *bioRxiv*, page 2022.09. 30.510297, 2022.

[Zhong *et al.*, 2024] Chengwei Zhong, Kok Siong Ang, and Jinmiao Chen. Interpretable spatially aware dimension reduction of spatial transcriptomics with stamp. *Nature Methods*, pages 1–12, 2024.

[Zhou *et al.*, 2025] Yuansheng Zhou, Xue Xiao, Lei Dong, Chen Tang, Guanghua Xiao, and Lin Xu. Cooperative integration of spatially resolved multi-omics data with cosmos. *Nature Communications*, 16(1):27, 2025.

[Zhu *et al.*, 2024a] James Zhu, Yunguan Wang, Woo Yong Chang, Alicia Malewska, Fabiana Napolitano, Jeffrey C Gahan, Nisha Unni, Min Zhao, Rongqing Yuan, and Fangjiang Wu. Mapping cellular interactions from spatially resolved transcriptomics data. *Nature methods*, 21(10):1830–1842, 2024.

[Zhu *et al.*, 2024b] Shijia Zhu, Naoto Kubota, Shidan Wang, Tao Wang, Guanghua Xiao, and Yujin Hoshida. Stie: Single-cell level deconvolution, convolution, and clustering in in situ capturing-based spatial transcriptomics. *Nature communications*, 15(1):7559, 2024.

[Zhu *et al.*, 2024c] Y. Zhu, X. He, C. Tang, X. Liu, Y. Liu, and K. He. Multi-view adaptive fusion network for spatially resolved transcriptomics data clustering. *IEEE Transactions on Knowledge and Data Engineering*, 36(12):8889–8900, 2024.