Appendix

Datasets and Preprocessing

We evaluate the proposed method on five widely used spatial transcriptomics datasets that include image information: (1) NanoString Lung 9-1, consisting of 20 fields of view (FOVs) with 960-dimensional gene expression data, synthetic 4-channel images (DAPI, PanCK, CD45, and CD3), and cell coordinate information. For this dataset, we cropped 120×120 pixel (21.6 μm × 21.6 μm) images for each cell and constructed a spatial graph with neighbors defined by a spatial distance of <80 pixels (14.4 μm). (2) 10x Visium DLPFC, comprising 10 tissue sections with the top 3000 highly variable genes identified using the Seurat pipeline, high-resolution H&E-stained images, and spatial coordinates. For this dataset, 50×50 pixel (38.7 μm × 38.7 μm) H&E-stained images were cropped, and neighbors were defined as spots within <150 pixels (116 μm). (3) 10x Visium human breast cancer data, containing 3000 highly variable genes, high-resolution H&E images, and spatial coordinates. For this dataset, 132×132 pixel H&E images were extracted, with neighbors defined as those within <400 pixels (116 μm). (4) Mouse brain anterior slice and (5) mouse brain coronal slice, both also including 3000 highly variable genes, high-resolution H&E images, and spatial coordinates. For these datasets, 30×30 pixel H&E images were cropped, and a spatial graph was constructed for spots within <188 pixels (116 μm).

Baseline Methods

Below are brief descriptions of the bseline methods compared to IE-HERCL:

1. **Scanpy**: Scanpy applies principal component analysis (PCA) to reduce the dimensionality of high-dimensional gene expression data. The reduced data is used to construct a neighborhood graph, followed by the Leiden algorithm to partition cells or spatial spots into distinct clusters.
2. **STAGATE**: STAGATE is a graph attention autoencoder framework that identifies spatial domains by integrating spatial coordinates and gene expression profiles. It learns neighborhood similarity using attention mechanisms, denoises data, and preserves spatial structure. It is also extensible to multi-slice data for 3D spatial domain extraction.
3. **GraphST**: GraphST is a graph-based self-supervised contrastive learning method that jointly encodes gene expression, spatial location, and local context through graph neural networks, learning informative and discriminative representations.
4. **stDCL**: stDCL is a dual-graph contrastive learning framework that integrates gene expression and spatial information via a graph autoencoder. It introduces dual contrastive learning strategies to enforce consistency between spatial structure and clustering patterns, enabling spatial domain identification and gene regulation interpretation.
5. **SiGra**: SiGra is a hybrid graph transformer method that integrates multichannel immunohistology images with spatial transcriptomics data. It constructs a hybrid graph transformer on a single-cell spatial graph, enhancing spatial domain detection and imputation accuracy in sparse expression data.
6. **xSiGra**: xSiGra is an interpretable graph-based framework that constructs spatial cellular graphs by incorporating immunohistology images and gene expression as node features. It uses a hybrid graph transformer along with an enhanced class activation mapping mechanism to reveal critical genes and cells for various cell types, enabling deeper biological interpretability from spatial data.

Evaluation Metrics

We use five metrics to evaluate spatial domain recognition performance: Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), Adjusted Mutual Information (AMI), Fowlkes-Mallows Index (FMI), and Homogeneity Score (HS):

1. ARI: Measures clustering similarity, ranging from -1 to 1. The formula is $\mathrm{ARI} = \dfrac{RI - E(RI)}{\max(RI) - E(RI)}$, where RI is the Rand Index and $E(RI)$ is its expected value.

2. NMI: Quantifies the shared information between clustering results, ranging from 0 to 1. The formula is $\mathrm{NMI}(X,Y)=\frac{I(X,Y)}{\sqrt{H(X)H(Y)}}$, where I($X$;$Y$) is mutual information between the predicted and true labels, and $H(X)$, $H(Y)$ are the entropies.

3. AMI: Adjusts mutual information for chance, making it more robust to variations in cluster size. The formula is $\mathrm{AMI}=\frac{I(X;Y)-E\left[I(X;Y)\right]}{\max\left(H(X),H(Y)\right)-E\left[I(X;Y)\right]}$, where $I(X;Y)$ is the mutual information and $E[I(X;Y)]$ is its expected value under random labeling.

4. FMI: Balances clustering precision and recall, ranging from 0 to 1. The formula is $\mathrm{FMI}=\sqrt{\frac{TP}{TP+FP}\times\frac{TP}{TP+FN}}$, where TP, FP, and FN represent true positives, false positives, and false negatives, respectively.

5. HS: Evaluates whether each cluster contains samples from a single category, ranging from 0 to 1, with higher values indicating better homogeneity.

Spatial Domain Identification via Clustering and Refinement

After training the model, we utilize the learned consensus representation to assign spots to different spatial domains using the non-spatial clustering method mclust [1], where each cluster is considered a spatial domain. The number of clusters is set to match the ground truth labels. For all datasets except the Nanostring Lung 9-1 dataset, we perform domain category correction. Specifically, for a given spot i, points within a radius r (default r=50) are regarded as its neighbors. Spot i is then reassigned to the domain corresponding to the most frequent label among its neighboring spots.

Experimental Results and In-depth Analysis

To comprehensively evaluate the effectiveness of the proposed method, we conduct spatial domain identification and detailed analysis across a variety of spatial transcriptomics datasets. The results demonstrate strong spatial coherence, high alignment with biological annotations, and superior performance in both quantitative metrics and qualitative assessments.

1. NanoString Lung9_Rep1 Dataset: We first perform spatial cell type visualization on all FOVs of the NanoString Lung9_Rep1 dataset (Figure 2). The proposed IE-HERCL method shows high consistency with the expert-annotated ground truth, particularly in accurately identifying the tumor spatial domain. In comparison to other methods, our model exhibits better boundary delineation between tumor and surrounding fibroblast or immune regions, which are often challenging to distinguish due to subtle transcriptional similarities in the tumor microenvironment. The overall ARI reaches 0.5679, confirming the effectiveness of the model in this complex setting.

2. DLPFC (Slice 151673) Dataset: On the DLPFC 151673 slice dataset, spatial domain visualization (Figure 3) reveals accurate recovery of anatomical structures, including white matter and layered cortical regions. In contrast, methods such as Scanpy, which rely solely on gene expression, often fail to distinguish white matter or merge adjacent layers. The learned latent representations (Figure 4) exhibit clear cluster separations in UMAP space, further supporting the model's capacity for discriminative embedding learning. In addition, the PAGE plot (Figure 5) reveals a smooth developmental trajectory across cortical layers, closely resembling the biological gradient observed in previous trajectory inference studies [2, 3]. This alignment supports the model's ability to capture continuous spatial transitions reflective of tissue development.

3. Human Breast Cancer Dataset: We further evaluate the model on the human breast cancer dataset (Figure 6). The clustering results reveal precise identification of biologically meaningful regions such as Health_1, Tumor_edge_2, and IDC_3, which are often ambiguously or partially detected by other methods. The ARI metric further confirms the strong performance, and the accurate delineation of transitional zones between healthy and malignant regions provides meaningful insights into tumor heterogeneity and progression, which may be beneficial for downstream pathological analysis.

4. Mouse Brain Anterior Dataset: Validation on the mouse anterior brain dataset is shown in Figure 7. The proposed method accurately identifies critical anatomical structures, such as the caudate putamen (CPu) and AON::L2 regions, with spatial domains matching those from the Allen Mouse Brain Reference Atlas. In contrast, other methods yield irregular or fragmented patterns in these areas. The ARI score indicates that our model outperforms the second-best baseline, stDCL, by 5.69%, demonstrating a stronger capacity for recovering biologically consistent spatial organization.

5. Mouse Coronal Brain Dataset: Finally, experiments on the mouse coronal brain dataset (Figure 8) show that IE-HERCL successfully identifies all seven annotated categories, with particularly accurate recognition of category 3, which is commonly overlooked by other methods. Furthermore, regions 2 and 3 are clearly separated, whereas competing methods tend to merge them. This finer granularity of spatial separation highlights the model's ability to resolve closely situated anatomical structures, aiding in the detailed mapping of functional and developmental organization in the mouse brain.

Across all datasets, the proposed IE-HERCL method demonstrates improved spatial coherence, finer granularity, and better alignment with known biological structures. These results not only confirm the model's robustness in handling diverse tissue types and complex spatial patterns, but also highlight its potential in facilitating downstream spatial transcriptomics analyses and biological interpretation.

Parameter Analysis

In addition, we analyze the influence of the smoothing parameter σ in the heat kernel function, which controls the sensitivity of edge weights to sample similarity and consequently affects the graph's connectivity. As shown in Figure 9, the performance across all evaluation metrics remains relatively stable over a wide range of σ values. Based on this observation, we fix σ= 0.1 in all experiments to ensure consistent graph construction without extensive parameter tuning.

Ablation studies

To further verify the effectiveness of the dual encoders (AE and GraphSAGE) and the cross-modal attention mechanism, we conduct additional experiments on the NanoString Lung9-1 FOV2 dataset. As detailed in Table 2, removing the AE or GraphSAGE encoder individually leads to a performance drop of 5.69% and 6.48% in ARI, respectively. Excluding the cross-modal attention module causes an even greater performance loss of 4.52%, underscoring its critical role in adaptive feature fusion. These results reinforce the importance of each module and demonstrate the model's ability to effectively integrate heterogeneous modalities across datasets.

We also evaluate the impact of different backbone networks for image feature extraction. As shown in Table 3, although EfficientNet_B0 achieves the highest ARI (0.6473), followed closely by VGG16 (0.6308) and ResNet50 (0.6279), the performance differences among these models are relatively small. This observation demonstrates the robustness of our framework with respect to the choice of image encoder. In the main experiments, we adopt ResNet50 as the default image backbone, considering its competitive performance and widespread use in the literature.

Cross-modal Attention Analysis

To verify the effectiveness of cross-modal attention, we visualize the attention weights assigned to each modality across different spatial domains, as shown in Figure 10. The results demonstrate that in regions such as IDC and DCIS/LCIS, the model assigns higher attention to the image modality, suggesting that morphological features play a dominant role in distinguishing these cancerous areas. In contrast, in regions like Healthy_1 and Tumor_edge_2, the gene expression modality receives more attention, indicating that transcriptional signals are more informative in these contexts. This adaptive weighting highlights the model’s ability to selectively integrate complementary information from both modalities based on region-specific characteristics. Such behavior confirms the utility of cross-modal attention in capturing heterogeneous spatial patterns and enhancing the overall accuracy of spatial domain identification. The varying modality contributions also enhance interpretability, revealing which data type is more influential in different tissue environments.

Conclusion

IE-HERCL excels in performance but faces computational challenges with ultra-large datasets, mainly due to the resource-intensive similarity graph construction. Future work will focus on optimizing this step (e.g., batch sampling) and extending the framework to multi-omics data (e.g., spatial proteomics) for more efficient and comprehensive analysis.

Table 1. Model performance under different combinations of parameters λ1, λ2, and λ3.

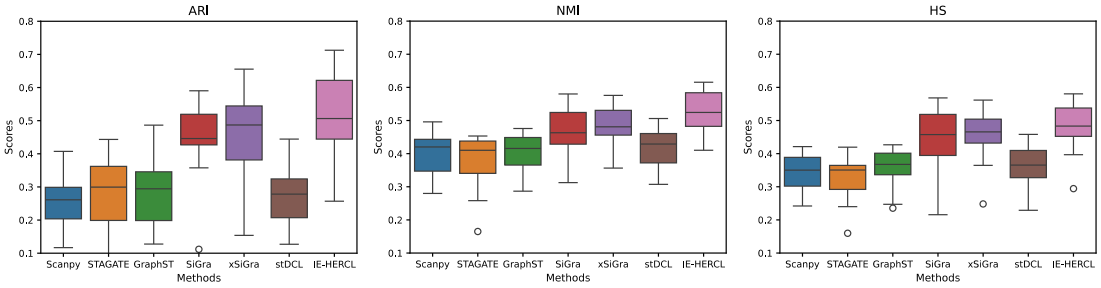
λ1	λ2	λ3	ARI	NMI	AMI	FMI	HS
1	0.0001	0.1	0.5562	0.6799	0.6740	0.5889	0.6730
1	0.0001	1	0.3981	0.5683	0.5635	0.4901	0.6706
1	0.001	0.1	0.5808	0.6861	0.6803	0.6117	0.6820
1	0.001	1	0.3600	0.5474	0.5412	0.4673	0.6611
10	0.0001	0.1	0.5688	0.6856	0.6800	0.6006	0.6836
10	0.0001	1	0.5388	0.6820	0.6763	0.5726	0.6749
10	0.001	0.1	0.5909	0.6790	0.6916	0.6211	0.6952
10	0.001	1	**0.6003**	**0.6949**	**0.6896**	**0.6325**	**0.7088**

Table 2. Ablation study on the NanoString Lung9-1 FOV2 dataset. This study evaluates the contributions of the AE, GraphSAGE, and cross-modal attention mechanisms. "w/o" denotes the exclusion of the corresponding component. The best results are highlighted in bold.

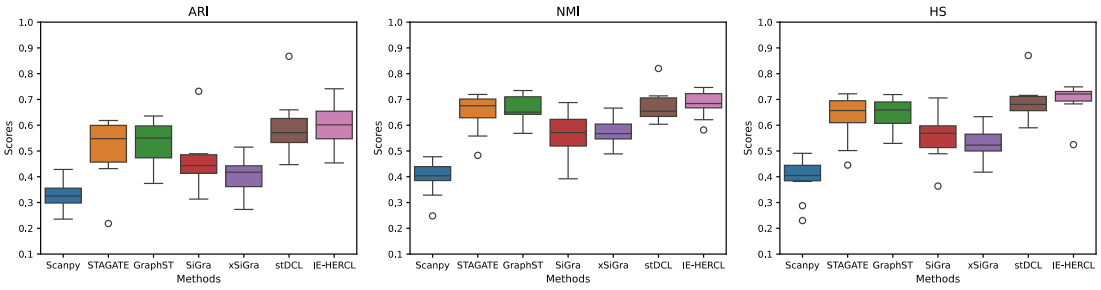
Methods	ARI	NMI	AMI	FMI	HS
IE-HERCL	**0.6970**	**0.5593**	**0.5578**	**0.8572**	**0.5172**
w/o AE	0.6401	0.5173	0.5151	0.8254	0.4510
w/o GraphSAGE	0.6322	0.5077	0.5054	0.8205	0.4369
w/o Attention	0.6518	0.5431	0.5410	0.8315	0.4720

Table 3. Impact of different image feature extraction methods on model performance on the human breast cancer dataset.

Methods	ARI	NMI	AMI	FMI	HS
ResNet34 [4]	0.6018	0.6783	0.6723	0.6343	0.6980
ResNet50 [5]	0.6279	**0.7133**	**0.7081**	**0.6757**	**0.7207**
ResNet101 [6]	0.6229	0.6950	0.6898	0.6552	0.7183
VGG16 [7]	0.6308	0.6988	0.6934	0.6605	0.7099
VGG19 [8]	0.6234	0.6856	0.6797	0.6529	0.6922
DenseNet121 [9]	0.6250	0.6978	0.6924	0.6576	0.7179
Efficientnet_B0 [10]	**0.6473**	0.7087	0.7035	0.6748	0.7194



(a) NanoString Lung 9-1



(b) DLPFC

Figure 1. Clustering performance of all methods on NanoString Lung 9-1 and DLPFC datasets.

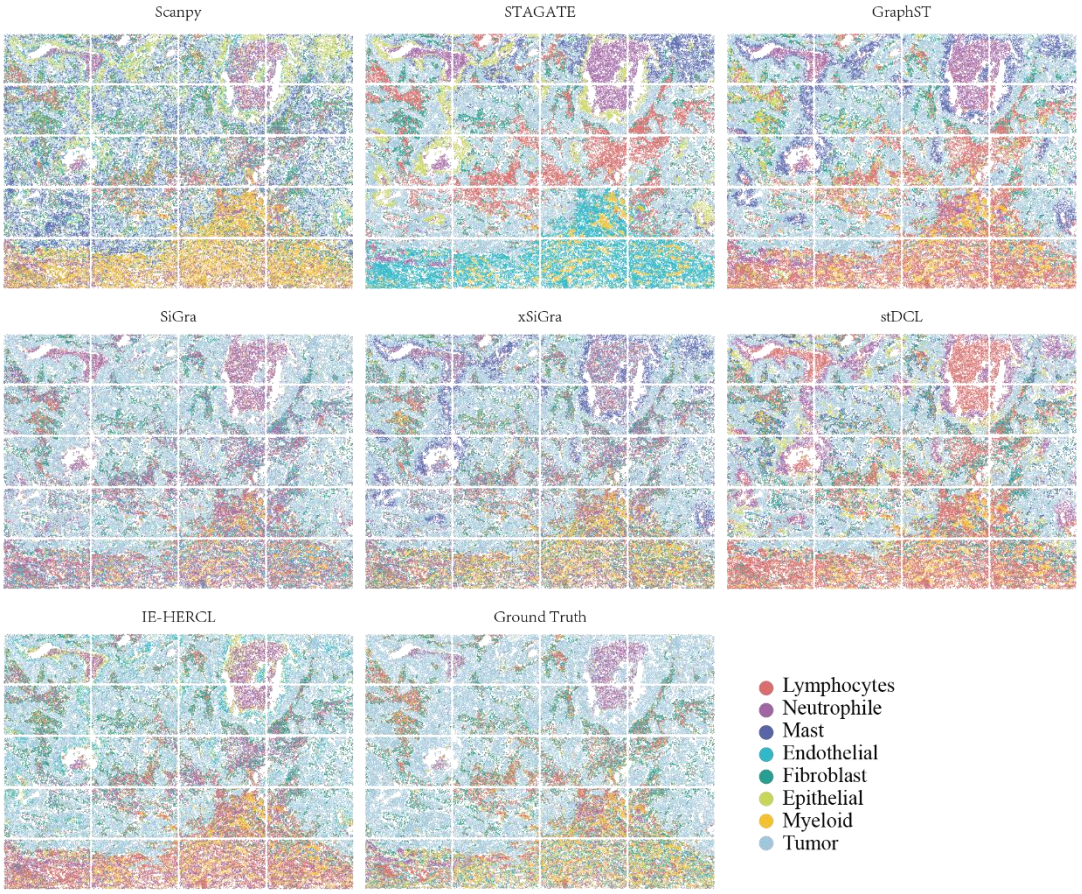


Figure 2. Spatial cell type identified by all methods on the NanoString Lung 9-1 dataset.

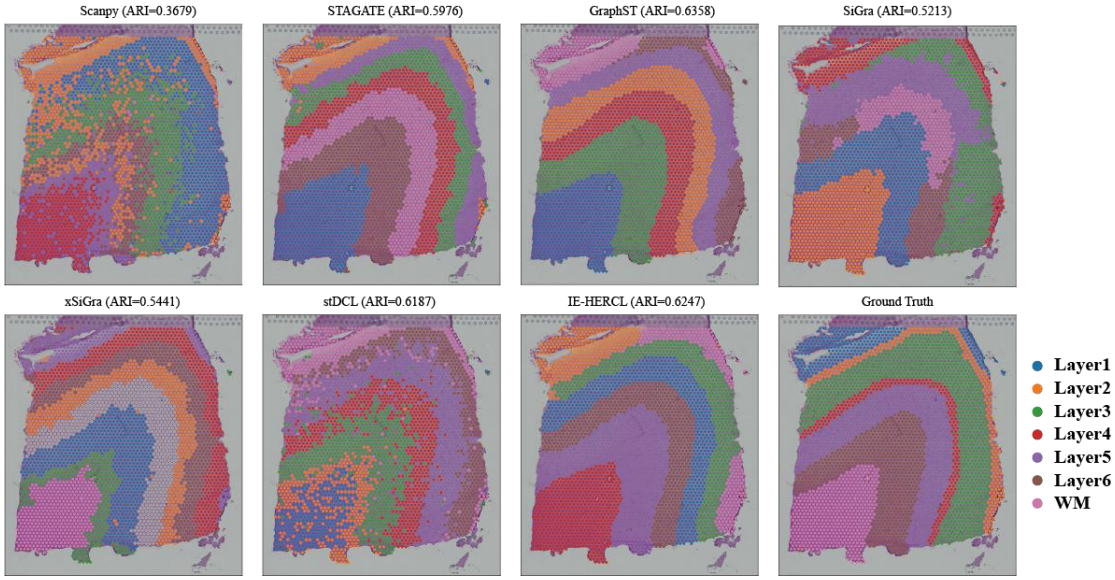


Figure 3. Spatial domains identified by all methods on the 151673 slice.

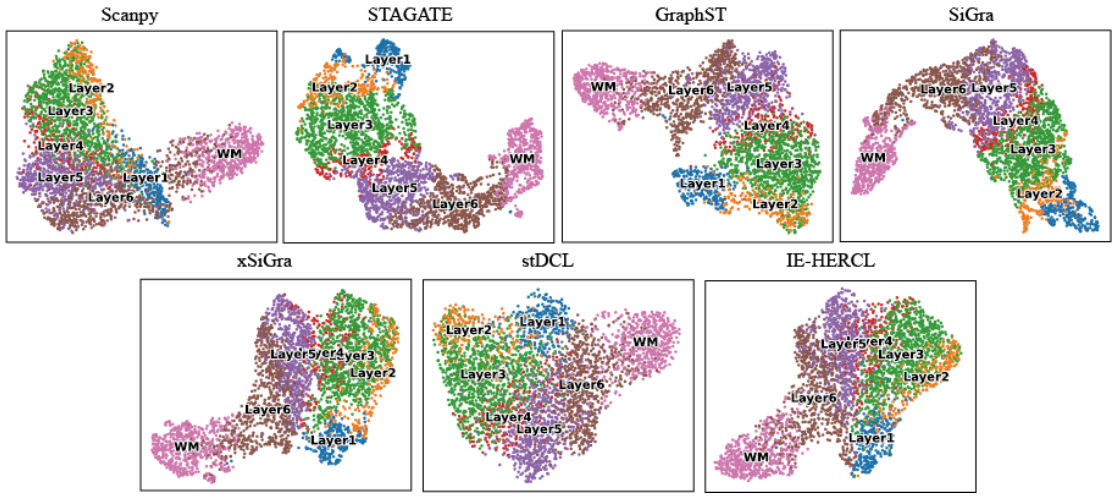


Figure 4. UMAP visualization of embeddings generated by all methods on the 151673 slice.

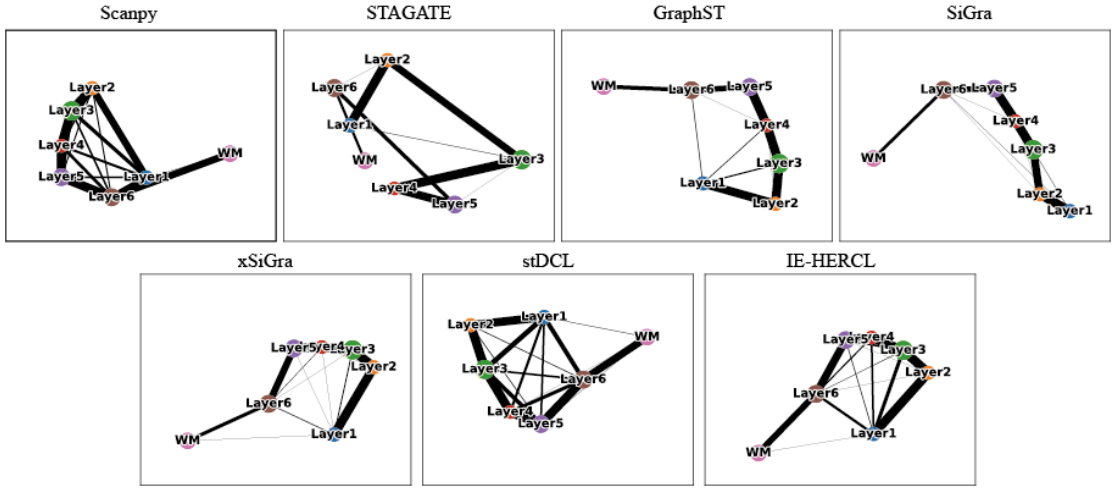


Figure 5. PAGA plot of embeddings generated by all methods on the 151673 slice.

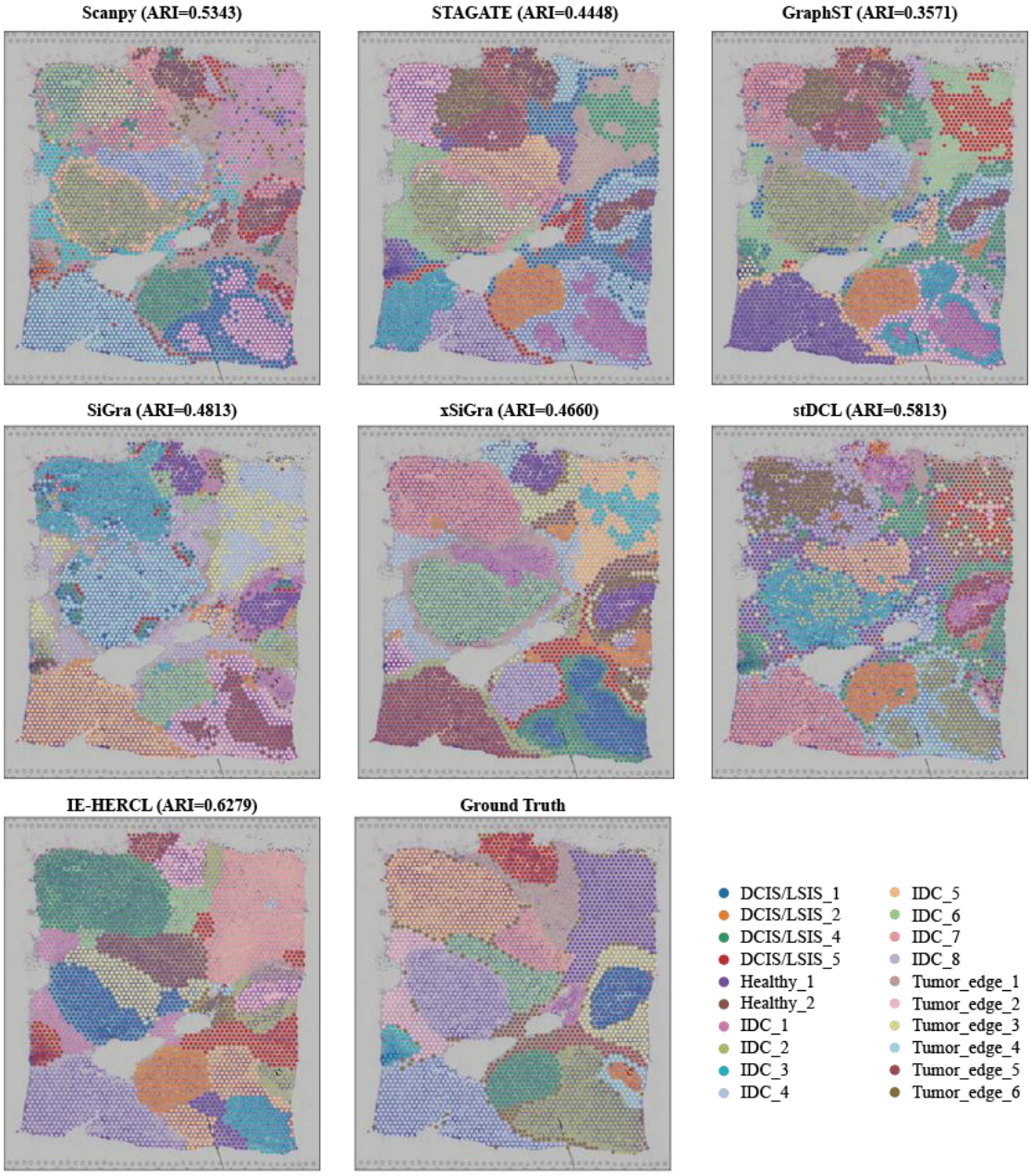


Figure 6. Spatial domains identified by all methods on the Human breast cancer dataset.

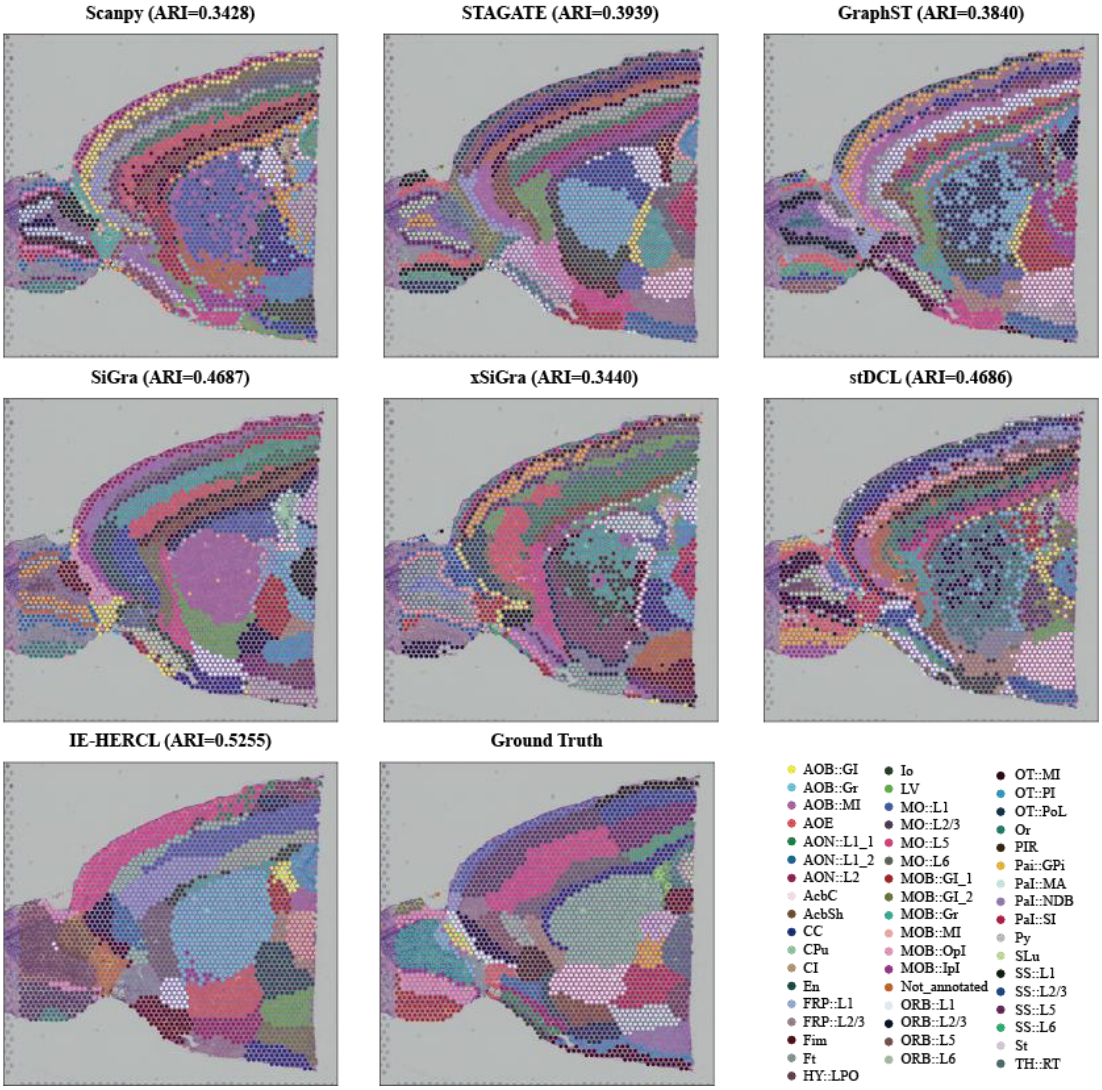


Figure 7. Spatial domains identified by all methods on the Mouse anterior brain dataset.

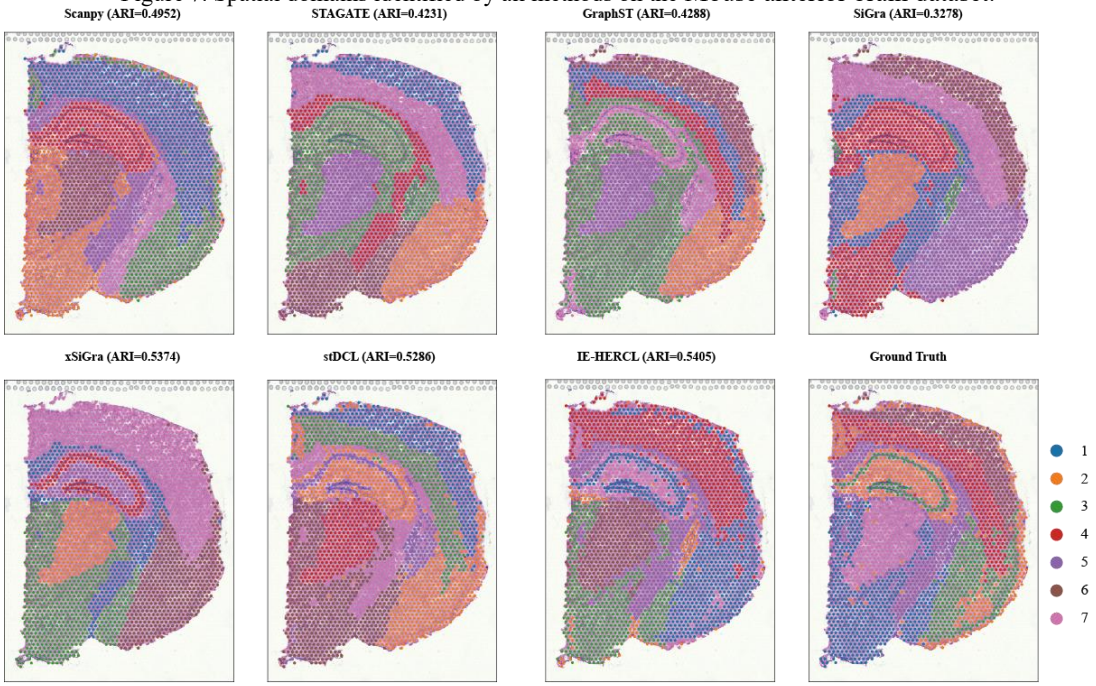


Figure 8. Spatial domains identified by all methods on the Mouse coronal brain dataset.

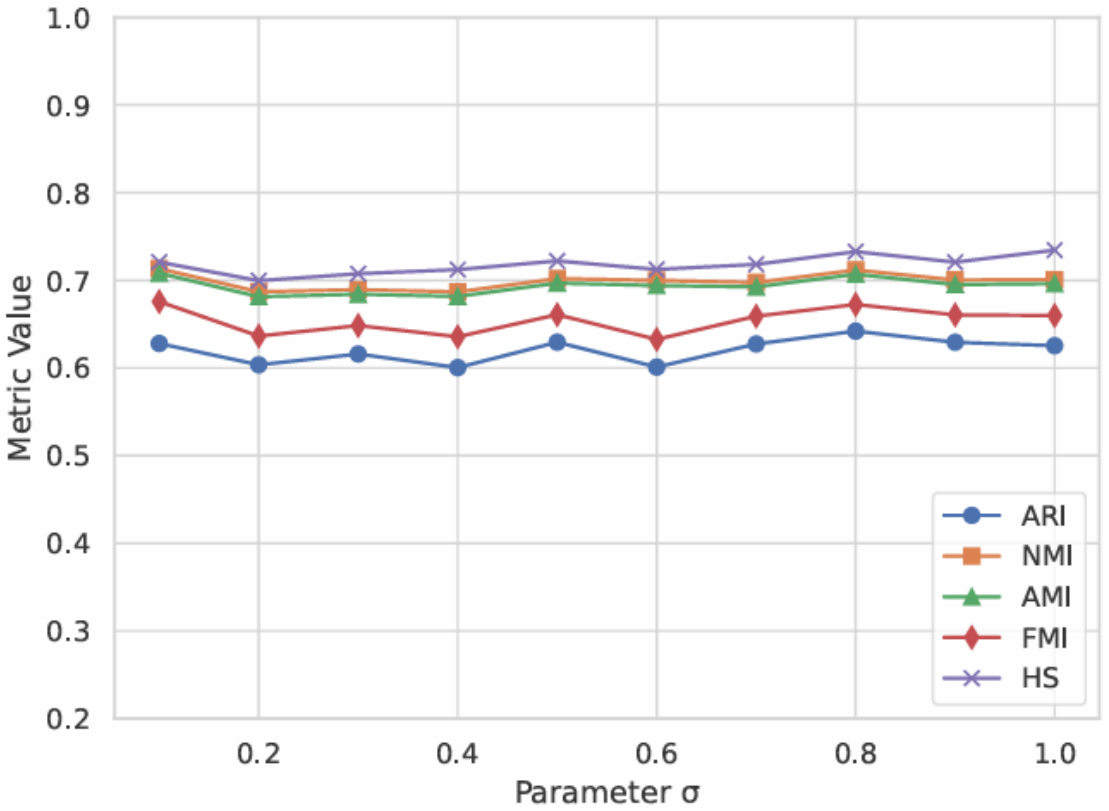


Figure 9. Performance of parameter σ on the human breast cancer dataset.

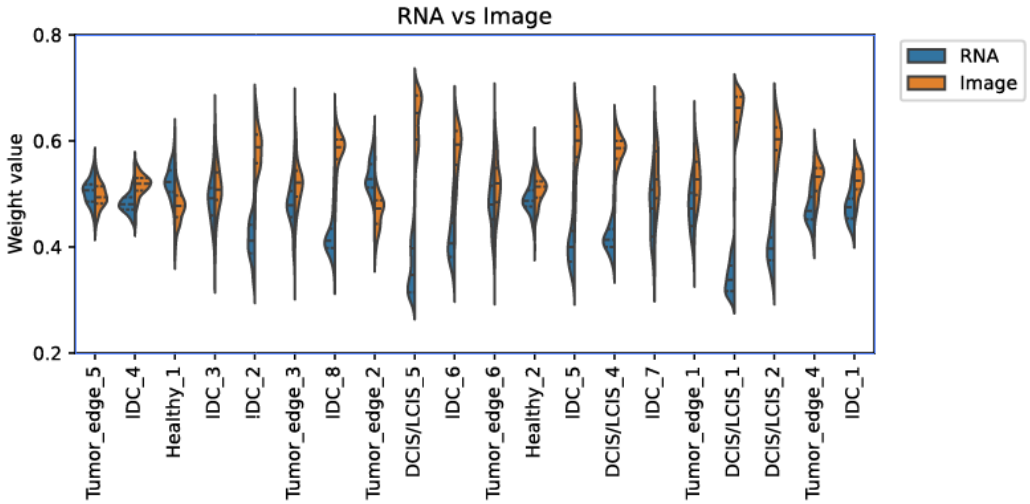


Figure 10: Cross-modality weights on the human breast cancer.

References

[1] Y. Long *et al.*, "Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with GraphST," *Nature Communications,* vol. 14, no. 1, p. 1155, 2023.

[2] R. R. Stickels *et al.*, "Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2," *Nature biotechnology,* vol. 39, no. 3, pp. 313-319, 2021.

[3] E. C. Gilmore and K. Herrup, "Cortical development: layers of complexity," *Current Biology,* vol. 7, no. 4, pp. R231-R234, 1997.

[4] Q. Zhuang, S. Gan, and L. Zhang, "Human-computer interaction based health diagnostics using ResNet34 for tongue image classification," *Computer Methods Programs in Biomedicine,* vol. 226, p. 107096, 2022.

[5] A. S. B. Reddy and D. S. Juliet, "Transfer learning with ResNet-50 for malaria cell-image classification," in *2019 International conference on communication and signal processing (ICCSP)*, 2019, pp. 0945-0949.

[6] Q. Zhang, "A novel ResNet101 model based on dense dilated convolution for image classification," *SN Applied Sciences,* vol. 4, pp. 1-13, 2022.

[7] Z.-P. Jiang, Y.-Y. Liu, Z.-E. Shao, and K.-W. Huang, "An improved VGG16 model for pneumonia image classification," *Applied Sciences,* vol. 11, no. 23, p. 11185, 2021.

[8] S. Mascarenhas and M. Agarwal, "A comparison between VGG16, VGG19 and ResNet50 architecture frameworks for Image Classification," in *2021 International conference on disruptive technologies for multi-disciplinary research and applications (CENTCON)*, 2021, vol. 1, pp. 96-99.

[9] A. Swaminathan, C. Varun, and S. Kalaivani, "Multiple plant leaf disease classification using densenet-121 architecture," *Journal of Electrical Engineering & Technology,* vol. 12, no. 5, pp. 38-57, 2021.

[10] M. UÇan, B. Kaya, and M. Kaya, "Multi-class gastrointestinal images classification using EfficientNet-B0 CNN model," in *2022 International Conference on Data Analytics for Business and Industry (ICDABI)*, 2022, pp. 1-5.