GE2Hist: Generating Histology Images from Single-cell Gene Expression via Cross-modal Generative Network

Hongmin Cai[1], Boan Ji[1], Shangyan Cai[2], Yi Liao[1], Jiazhou Chen[3], and Weitian Huang[1,4,*]

[1] School of Computer Science and Engineering, South China University of Technology, Guangzhou, China
[2] Department of Computer Science, Hong Kong Baptist University, Hong Kong
[3] School of Computer Science and Technology, Guangdong University of Technology, Guangzhou, China
[4] Guangdong Institute of Intelligence Science and Technology, Zhuhai, China

Abstract. Histological images are essential in biomedical research and diagnosis, extending beyond detailed cell and tissue morphology to provide an intuitive view of the cellular microenvironment and spatial relationships. While single-cell gene expression data reveal molecular distinctions in cell states, their complexity obscures cellular interactions and spatial organization. To overcome this, reconstructing histological images from large-scale single-cell data is essential for intuitively visualizing spatial architecture. This paper proposes a single-cell-level histological image generation method that derives cell state representations from gene expression data using a single-cell foundation model. A conditional diffusion model is leveraged to generate histological images, accurately reconstructing the cellular microenvironment and spatial cell type distribution. By decoupling cellular state into two components, cell type and microenvironment, we propose two complementary approaches for generating pathology images, one conditioned on scRNA-seq data and the other on cell type. Our approach successfully generates high-quality histological images of human breast and colon cancer tissues, capturing key spatial features such as cell density, compositional distribution, and cell spacing within tissues.

Keywords: Cross-modal generation · Histology image · Diffusion model.

1 Introduction

The rapid progress in single-cell genomics has deepened our understanding of cellular heterogeneity and tissue microenvironments. However, traditional scRNA-seq techniques lose crucial spatial information when cells are dissociated from their native tissue contexts, hampering our ability to fully grasp cell-cell interactions and tissue dynamics [9, 23]. Spatial transcriptomics has emerged as a powerful solution, preserving spatial context by measuring gene expression in situ and revealing the complexity of cellular niches [10, 11, 14, 15, 22, 24, 25].

A niche is the local microenvironment of a cell within a tissue, encompassing its spatial location, surrounding cell types, and intercellular interactions. These elements collectively shape cellular behavior and function. Studying niches is crucial as it helps elucidate how cells are influenced by their surroundings. It reveals the spatial heterogeneity of tissues and the dynamic interactions that drive development, homeostasis, and disease. This understanding is vital for advancing regenerative medicine, improving disease diagnostics, and developing targeted therapies. Histological imaging offers the most intuitive way to observe cellular niches, and multimodal studies using paired histological images and gene expression data have demonstrated significant potential in disease diagnosis [1, 4, 5, 7, 8] and cellular composition analysis [13]. However, the limited accessibility of spatial transcriptomics technology poses a challenge in obtaining paired histological images and single-cell gene expression data. Consequently, reconstructing histological images of cellular niches from gene expression data holds significant potential for various applications.

Building on this, recent efforts to integrate single-cell and spatial data have shown promise in inferring spatial features from gene expression. For example, the Nicheformer model successfully demonstrated the feasibility of predicting cellular niches from scRNA-seq data [20]. Yet, this method primarily yields numerical results, lacking the intuitive visual insights provided by histological images, which are essential for understanding cellular niches. Previous efforts, such as RNA-GAN [2] and RNA-CDM [3], relied on bulk-level RNA sequencing datasets from resources like the Genotype-Tissue Expression (GTEx) project and the Cancer Genome Atlas (TCGA). These bulk datasets provide average gene expression profiles across entire tissue samples, which can only capture overall tissue characteristics and fail to represent the heterogeneity of individual cells.

In contrast, this study introduces a novel method to generate histological images of cellular niches from single-cell gene expression data obtained through advanced preprocessing of subcellular-level spatial transcriptomic data from the Visium HD library. Using a single-cell foundation model, we extract and decouple cellular state information into cell type and environmental context. This enables the development of two image-generation pathways via a conditional diffusion model: one based on scRNA-seq data and the other on specific cell types. The resulting images reflect gene expression patterns and reveal cell density and neighborhood composition, providing an intuitive tool for studying the cellular microenvironment. As an exploratory advancement in the field of cellular digital twins, our method can precisely visualize how interventions, such as simulated gene editing or drug perturbations, might reshape this cellular ecosystem. This provides critical visual evidence for predicting cellular responses, holding the potential to accelerate therapeutic development and reduce development costs.

The main contributions of this work are as follows:

1. A novel algorithm for generating histological images from single-cell gene expression data has been developed, utilizing a conditional diffusion model

to reconstruct cell morphology and surrounding environment from cellular state information.
2. Cellular states in the latent space have been decomposed into two independent variables, enabling the implementation of two distinct generation methods to produce more diverse histological images on demand.
3. The proposed model produces the highest-quality images compared to other approaches and demonstrates strong consistency with real spatial images, particularly in terms of cell density, compositional distribution, and cell spacing.

2 Method

We present GE2Hist, a model designed to generate histological images X from single-cell gene expression data Y. The low-dimensional embedding E represents the gene expression processed through the foundation model. $\{X_0, \ldots, X_T\}$ represent the noise images within the diffusion generation process, where X_0 specifically refers to the final synthesized histological image. Since cell states are determined by cell types and their microenvironments, we introduce z_c as the variable that characterizes cell types, z_e as the variable that represents the microenvironment of the cells, and z_s as the variable that encapsulates the overall cellular state.

2.1 scRNA-seq Embedding & Decoupling

The state of a cell is influenced by both its cell type and the surrounding microenvironment. To create more effective generative control conditions, we use scGPT [6]—a single-cell foundation model trained on 33 million cells—to encode RNA-seq data into a low-dimensional vector E. This process removes biases from measurement techniques and external factors, enhancing the extraction of biological features. Subsequently, two Variational Autoencoders are used to decouple E into two independent latent variables representing cell type and microenvironment: $q(z_c|E) = \mathcal{N}(z_c; \mu_c(E), \sigma_c^2(E))$, $q(z_e|E) = \mathcal{N}(z_e; \mu_e(E), \sigma_e^2(E))$, where $\mu_c, \sigma_c^2 = Encoder_c(E; \phi_c)$ and $\mu_e, \sigma_e^2 = Encoder_e(E; \phi_e)$

Then, to ensure proper decoupling of the two components, we applied distinct constraints to z_e and z_c. To guide the accurate encoding of cell type information in z_c, we set a classifier using cross-entropy loss to categorize z_c.

$$\mathcal{L}_{\text{cls}}(\phi_c, \theta_{cls}) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} y_{ij} \log(\hat{y}_{ij}) \tag{1}$$

Where N is the number of samples, C is the number of classes, y_{ij} is the true label (one-hot encoded from scGPT), and $\hat{y}_{ij} = Classifier(Encoder_c(E_i; \phi_c); \theta_{cls})_j$. Additionally, we used KL divergence as a regularization term for z_e, ensuring a well-structured and interpretable latent representation that supports stable training and effective data generation:

$$\mathcal{L}_{\text{KL}}(\phi_e) = D_{KL}(q(z_e|E)||p(z_e)). \tag{2}$$

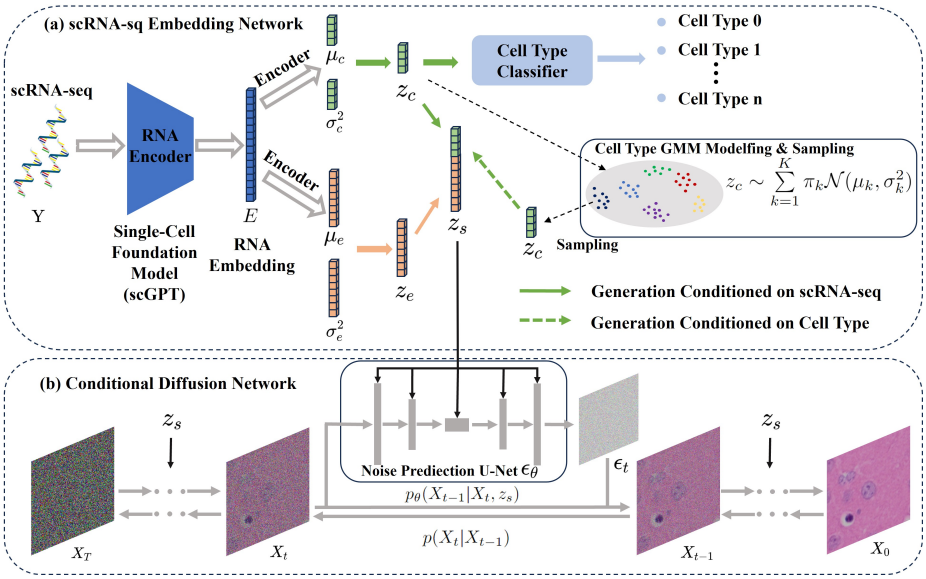

Fig. 1. Overview of the GE2Hist model for generating histological images from single-cell gene expression data. (a) Gene expression data from single-cell RNA sequencing is embedded into a low-dimensional latent space using a foundation model. This representation is then decoupled into cell type and microenvironment information. (b) Conditional diffusion generation model for synthesizing histological images from cell states.

Then, z_c and z_e will be merged into z_s to indicate the cell's state: $z_s = concat(z_c, z_e)$.

With the classifier's guidance, z_c is naturally clustered by cell type in the feature space. We fit this distribution using a Gaussian mixture model. Based on this Gaussian mixture model, we can sample z_c from specific clusters, which forms the basis for our generation conditioned on cell type. The prior distribution $p(z_e)$ for the latent microenvironment variable z_e is assumed to be a standard normal distribution $N(0, 1)$.

2.2 Conditional Diffusion

Diffusion models generate images by starting with random Gaussian noise and progressively transforming it into a coherent image. This stochastic nature leads to diverse outputs, but the generated images are often uncontrollable and may not match our desired content. To tackle this challenge, many studies have proposed effective strategies for controlling the image generation process [13, 16–19]. In this work, we focus on using cellular states to guide the diffusion model in generating histological images.

This paper presents two methods for obtaining representations of cell types, each tailored for different generation objectives:

1. Generation conditioned on scRNA-seq: The vector z_c is obtained directly from the gene expression embedding of the cells, as illustrated by the solid arrows in Fig. 1. This generation method is primarily used to restore cellular morphology and the surrounding environment from gene expression data.
2. Generation conditioned on cell type: The vector z_c is obtained by sampling from the clusters corresponding to the specific cell type, as illustrated by the dashed arrows in Fig. 1. This generation method enables independent observation of the effects of cell type and environment on cellular morphology, facilitating a deeper understanding of cellular functions and interactions.

Equation (2) shows two different approaches,

$$z_c = \begin{cases} f_\Phi(E) & \text{Generation conditioned on scRNA-seq} \\ z_c \sim \mathcal{N}\left(\mu_c, \sigma_c^2 \mathbf{I}\right) & \text{Generation conditioned on cell type} \end{cases} \tag{3}$$

Introducing control conditions does not affect the model's forward diffusion, which can be expressed in terms of conditional probability as follows,

$$p(X_T \mid X_0) = \int \prod_{t=1}^{T} \mathcal{N}(X_t; \sqrt{\alpha_t} X_{t-1}, (1-\alpha_t)\mathbf{I})\, dX_1 \cdots dX_{T-1} \tag{4}$$

Where α_t is a pre-defined hyperparameter that controls the amount of noise added at each diffusion step t.

The reverse generation process of the conditional diffusion model involves modeling $p_\theta(X_{t-1}|X_t, z_s)$, and the reverse generation process can be described as follows,

$$p_\theta(X_1, X_2, ..., X_T \mid z_s) = p(X_T) \prod_{t=1}^{T} p_\theta(X_{t-1}|X_t, z_s) \tag{5}$$

The $p(X_T)$ represents the distribution of the initial noise, which is typically assumed to be a standard normal distribution. The product notation describes the process of gradually generating the image X_0 from the noise X_T, $p_\theta(X_{t-1}|X_t, z_s)$ can be expressed as follows,

$$p_\theta(X_{t-1}|X_t, z_s) = \mathcal{N}\big(X_{t-1}; \mu_\theta(X_t, t, z_s), \Sigma_\theta(X_t, t, z_s)\big). \tag{6}$$

In the conditional diffusion model, $\mu_\theta(X_t, t, z_s)$ represents the mean of X_{t-1} given the current noise level X_t, the conditional information z_s, and the time step t. This mean is typically parameterized by a neural network ϵ_θ, commonly called the noise prediction network.

$$\mu_\theta(X_t, t, z_s) = \frac{1}{\sqrt{\alpha_t}}\Big(X_t - \sqrt{1-\alpha_t}\epsilon_\theta(X_t, t, z_s)\Big). \tag{7}$$

This noise prediction network ϵ_θ is typically implemented using a UNet architecture. During model training, the optimization is achieved by minimizing

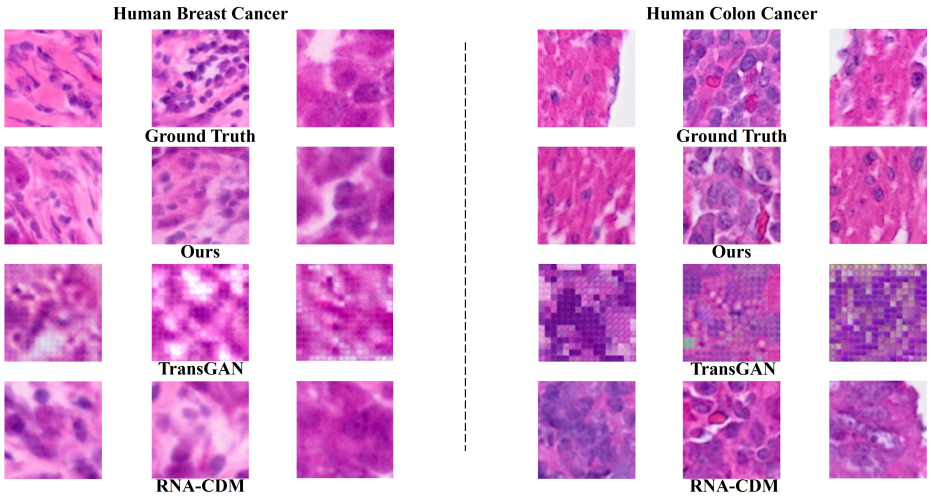

Fig. 2. Histology image generation conditioned on scRNA-seq.GE2Hist model accurately captured cell density, spatial composition, and cell type proportions in the generated histological images, showing high consistency with the ground truth.

the distance between the predicted noise ϵ_θ and the actual noise ϵ_t. The loss function of our optimization objective $\mathcal{L}_{\text{CDM}}$ can be expressed as follows,

$$\mathcal{L}_{\text{CDM}} = \sum_{t=1}^{T} \mathbb{E}_{X_0,\epsilon_t}\Big[\|\epsilon_\theta(X_0, t, z_s) - \epsilon_t\|_2^2\Big]. \tag{8}$$

The total loss function $\mathcal{L}_{\text{total}}$ for the GE2Hist model can be derived from equations (1)(2)(8),

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CDM}} + \beta\mathcal{L}_{\text{cls}} + \gamma\mathcal{L}_{\text{KL}}. \tag{9}$$

Where β and γ are the balancing coefficients to reconcile the three components, the KL divergence constraint regularizes the latent space to follow a normal distribution, ensuring a well-structured and interpretable latent representation that supports stable training and effective data generation. The default values for β and γ are set to 0.2 and 0.1.

3 Experiments

High-dimensional (>19k) RNA sequences are encoded into a structured 64D latent vector, $z_s = [z_c, z_e]$, first via a foundation model for compression to 256D, and then through a pre-trained VAE. Critically, the semantic decoupling of z_c and z_e is not pre-learned but is instead enforced during the main training phase of our diffusion model. This is achieved by jointly training the diffusion U-Net

Table 1. Comparison of KID and LPIPS

Methods	Data	KID	LPIPS
TransGAN	Breast Cancer	0.3970 ± 0.0179	0.3331 ± 0.0350
	Colon Cancer	0.5386 ± 0.0106	0.3731 ± 0.0345
RNA-CDM	Breast Cancer	0.0466 ± 0.0126	0.1641 ± 0.0574
	Colon Cancer	0.0576 ± 0.0147	0.1740 ± 0.0686
ours	Breast Cancer	**0.0164 ± 0.0137**	**0.1538 ± 0.0626**
	Colon Cancer	**0.0286 ± 0.0156**	**0.1657 ± 0.0667**

and a randomly initialized MLP classifier under a composite loss function that includes cell-type classification, KL divergence, and denoising terms. Our U-Net, which leverages ResNet blocks and cross-attention to condition on z_s, is trained with the Adam optimizer over 1000 diffusion steps using a cosine schedule.

3.1 Data Processing

In this study, we utilized human breast cancer and colon cancer samples from the Visium HD Spatial Gene Expression Library. Visium HD provides high-resolution H&E-stained images and gene expression data through high-density barcode arrays, enabling precise localization of gene expression within tissue sections [21]. Image segmentation was used to create nuclear masks, which were then used for binning to assign barcodes to individual nuclei. After spatial verification to filter out overlapping barcodes, UMI counts were summed to generate a single-cell spatial expression matrix, reflecting the gene expression profiles of individual cells.We used a foundation model to mitigate the highly imbalanced cell counts across 9 types in breast cancer and 38 in colon cancer, with cell labels provided by a pre-trained model.

3.2 Results

We conducted image-generation experiments using human breast and colon cancer tissues based on scRNA-seq data (Fig. 2). Our GE2Hist model accurately captured cell density, spatial composition, and cell type proportions in the generated histological images, showing high consistency with the ground truth. Our method infers the surrounding environment from the gene expression of a single central cell, so while generated images may not match the ground truth exactly, key metrics like cell density and spatial composition are well-preserved, indicating accurate niche inference. Unlike TransGAN [12], which produced low-quality and non-diverse images, and RNA-CDM [3], which showed discrepancies in cellular niche structures, GE2Hist preserved key metrics and reflected cellular states and microenvironments, highlighting its superior performance. We quantitatively assessed our method's performance using Kernel Inception Distance(KID) and Learned Perceptual Image Patch Similarity(LPIPS) metrics, with results in Table 1.

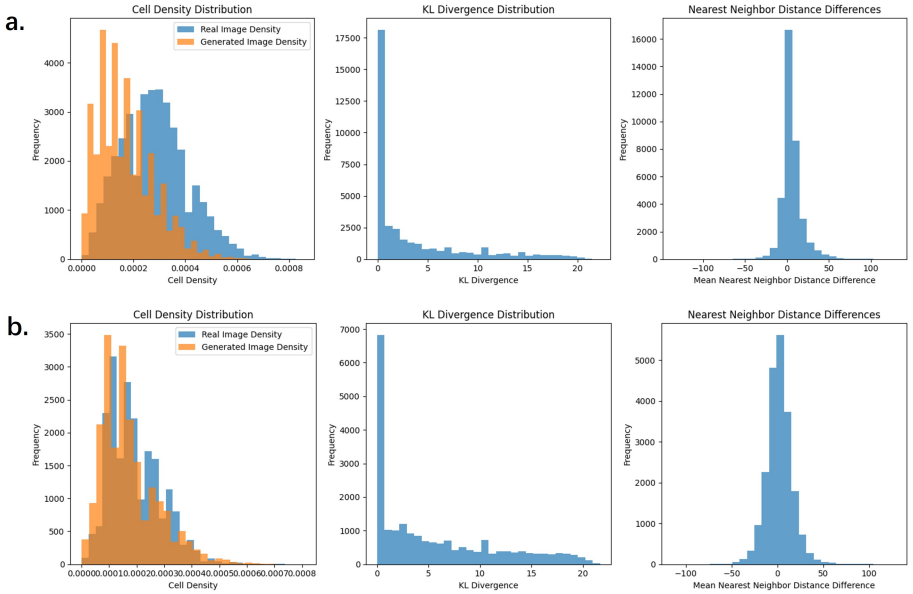

Fig. 3. GE2Hist accurately predicts neighborhood compositions for breast cancer(a) and colon cancer(b) cells. The left panel demonstrates that the cell density in the generated images closely matches the real images. The middle panel shows the KL divergence between the cell type distributions of the generated and real images. The right panel presents the mean nearest-neighbor distance differences in cell spacing.

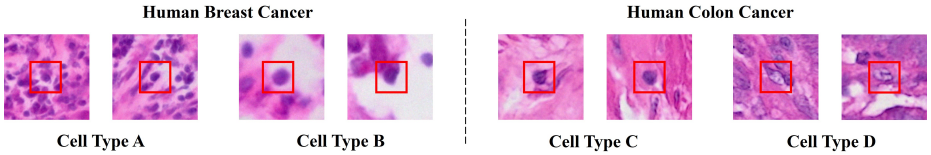

Fig. 4. Histology image generation conditioned on cell type. By sampling z_c, GE2Hist can generate histological images with specified central cell types.

To further validate the accuracy of GE2Hist in predicting cellular niches, we quantitatively assessed cell density, cell type distribution differences, and cell spacing distribution differences(Fig 3). Cell density was determined by counting the number of cells per unit area in each image. Cell type distribution differences were evaluated by calculating the proportions of different cell types and comparing the KL divergence between the generated and real distributions, with values close to zero indicating consistency. Cell spacing distribution differences were assessed by computing the nearest neighbor distances among cells, calculating their mean and variance, and comparing these metrics between the generated and real images, where values close to zero suggest spatial consistency.

We evaluated the cell type classifier for z_c in our model(Breast cancer classification accuracy: 92.31%; colon cancer: 87.93%). Breast cancer tissues had 9 cell types, while colon cancer tissues had 38, making the latter's classification metrics slightly lower. Guided by the classifier, we calculated the Gaussian mixture distribution for z_c and sampled from specific clusters to generate histological images with specific cell types (Fig. 4).

4 Conclusion

This paper introduces a novel method to generate histological images of cellular niches from single-cell gene expression data, leveraging a single-cell foundation model and a conditional diffusion model. Through the decoupling of cellular states, our method enables histological image generation from scRNA-seq data and supports the generation of images with specified cell types. The generated images exhibit high consistency with real images in terms of cell density, cell type proportions, and cell spacing distribution, accurately depicting the spatial context between cells. By enabling tissue reconstruction at single-cell resolution, our approach allows pathologists to explore how cells respond to environmental signals, promote tissue development, maintain homeostasis, and adapt to disease states from a morphological perspective.

Acknowledgments. This work was supported in part by the National Key Research and Development Program of China (2024YFF1206600), the National Natural Science Foundation of China (U21A20520, 62325204) and the Key-Area Research and Development Program of Guangzhou City (2023B01J1001).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Carrillo-Perez, F., Morales, J.C., Castillo-Secilla, D., Gevaert, O., Rojas, I., Herrera, L.J.: Machine-learning-based late fusion on multi-omics and multi-scale data for non-small-cell lung cancer diagnosis. Journal of Personalized Medicine **12**(4), 601 (2022)
2. Carrillo-Perez, F., Pizurica, M., Ozawa, M.G., Vogel, H., West, R.B., Kong, C.S., Herrera, L.J., Shen, J., Gevaert, O.: Synthetic whole-slide image tile generation with gene expression profile-infused deep generative models. Cell Reports Methods **3**(8) (2023)
3. Carrillo-Perez, F., Pizurica, M., Zheng, Y., Nandi, T.N., Madduri, R., Shen, J., Gevaert, O.: Generation of synthetic whole-slide image tiles of tumours from rna-sequencing data via cascaded diffusion models. Nature Biomedical Engineering pp. 1–13 (2024)
4. Castillo, D., Gálvez, J.M., Herrera, L.J., Román, B.S., Rojas, F., Rojas, I.: Integration of rna-seq data with heterogeneous microarray data for breast cancer profiling. BMC Bioinformatics **18**, 1–15 (2017)

5. Chen, R.J., Lu, M.Y., Williamson, D.F., Chen, T.Y., Lipkova, J., Noor, Z., Shaban, M., Shady, M., Williams, M., Joo, B., et al.: Pan-cancer integrative histology-genomic analysis via multimodal deep learning. Cancer Cell **40**(8), 865–878 (2022)
6. Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., Wang, B.: scgpt: toward building a foundation model for single-cell multi-omics using generative ai. Nature Methods **21**(8), 1470–1480 (2024)
7. Dawood, M., Branson, K., Rajpoot, N.M., Minhas, F.: Albrt: Cellular composition prediction in routine histology images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 664–673 (2021)
8. Diao, J.A., Wang, J.K., Chui, W.F., Mountain, V., Gullapally, S.C., Srinivasan, R., Mitchell, R.N., Glass, B., Hoffman, S., Rao, S.K., et al.: Human-interpretable image features derived from densely mapped cancer pathology slides predict diverse molecular phenotypes. Nature Communications **12**(1), 1613 (2021)
9. Du, J., Yang, Y.C., An, Z.J., Zhang, M.H., Fu, X.H., Huang, Z.F., Yuan, Y., Hou, J.: Advances in spatial transcriptomics and related data analysis strategies. Journal of Translational Medicine **21**(1), 330 (2023)
10. Fischer, D.S., Schaar, A.C., Theis, F.J.: Modeling intercellular communication in tissues using spatial graphs of cells. Nature Biotechnology **41**(3), 332–336 (2023)
11. He, S., Bhatt, R., Brown, C., Brown, E.A., Buhr, D.L., Chantranuvatana, K., Danaher, P., Dunaway, D., Garrison, R.G., Geiss, G., et al.: High-plex imaging of rna and proteins at subcellular resolution in fixed tissue by spatial molecular imaging. Nature Biotechnology **40**(12), 1794–1806 (2022)
12. Jiang, Y., Chang, S., Wang, Z.: Transgan: Two pure transformers can make one strong gan, and that can scale up. Advances in Neural Information Processing Systems **34**, 14745–14758 (2021)
13. Kohane, I.S., Churchill, S., Tan, A.L.M., Vella, M., Perry, C.L.: The digital–physical divide for pathology research. The Lancet Digital Health **5**(12), e859–e861 (2023)
14. Lu, Y., Liu, M., Yang, J., Weissman, S.M., Pan, X., Katz, S.G., Wang, S.: Spatial transcriptome profiling by merfish reveals fetal liver hematopoietic stem cell niche architecture. Cell Discovery **7**(1), 47 (2021)
15. Marx, V.: Method of the year: spatially resolved transcriptomics. Nature Methods **18**(1), 9–14 (2021)
16. Palumbo, E., Manduchi, L., Laguna, S., Chopard, D., Vogt, J.E.: Deep generative clustering with multimodal diffusion variational autoencoders. In: The Twelfth International Conference on Learning Representations (2024)
17. Pandey, K., Mukherjee, A., Rai, P., Kumar, A.: Diffusevae: Efficient, controllable and high-fidelity generation from low-dimensional latents. arXiv preprint arXiv:2201.00308 (2022)
18. Preechakul, K., Chatthee, N., Wizadwongsa, S., Suwajanakorn, S.: Diffusion autoencoders: Toward a meaningful and decodable representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10619–10629 (2022)
19. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)
20. Schaar, A.C., Tejada-Lapuerta, A., Palla, G., Gutgesell, R., Halle, L., Minaeva, M., Vornholz, L., Dony, L., Drummer, F., Bahrami, M., et al.: Nicheformer: a foundation model for single-cell and spatial omics. bioRxiv pp. 2024–04 (2024)
21. Sun, C., Zhang, Y.: Sthd: probabilistic cell typing of single spots in whole transcriptome spatial data with high definition. bioRxiv pp. 2024–06 (2024)

22. Varrone, M., Tavernari, D., Santamaria-Martínez, A., Walsh, L.A., Ciriello, G.: Cellcharter reveals spatial cell niches associated with tissue remodeling and cell plasticity. Nature Genetics **56**(1), 74–84 (2024)
23. Williams, C.G., Lee, H.J., Asatsuma, T., Vento-Tormo, R., Haque, A.: An introduction to spatial transcriptomics for biomedical research. Genome Medicine **14**(1), 68 (2022)
24. Yao, Z., Liu, H., Xie, F., Fischer, S., Adkins, R.S., Aldridge, A.I., Ament, S.A., Bartlett, A., Behrens, M.M., Van den Berge, K., et al.: A transcriptomic and epigenomic cell atlas of the mouse primary motor cortex. Nature **598**(7879), 103–110 (2021)
25. Yao, Z., van Velthoven, C.T., Kunst, M., Zhang, M., McMillen, D., Lee, C., Jung, W., Goldy, J., Abdelhak, A., Aitken, M., et al.: A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain. Nature **624**(7991), 317–332 (2023)