



Inferring Super-Resolved Gene Expression by Integrating Histology Images and Spatial Transcriptomics with HISTEX

Shuailin Xue¹, Changmiao Wang², Xiaomao Fan³, and Wenwen Min¹

¹School of Information Science and Engineering, Yunnan University, Kunming, China

²Medical Big Data, Shenzhen Research Institute of Big Data, Shenzhen, China

³College of Big Data and Internet, Shenzhen Technology University, Shenzhen, China

Correspondence author (✉): minwenwen@ynu.edu.cn

Abstract. The groundbreaking development of spatial transcriptomics (ST) enables researchers to map gene expression across tissues with spatial precision. However, current next-generation sequencing methods, which theoretically cover the entire transcriptome, face limitations in resolving spatial gene expression at high resolution. The recently introduced Visium HD technology offers a balance between sequencing depth and spatial resolution, but its complex sample preparation and high cost limit its widespread adoption. To address these challenges, we introduce HISTEX, a multimodal fusion approach that leverages a bidirectional cross-attention mechanism and a general-purpose foundation model. HISTEX integrates spot-based ST data with histology images to predict super-resolution (SR) spatial gene expression. Experimental evaluations demonstrate that HISTEX outperforms state-of-the-art methods in accurately predicting SR gene expression across diverse datasets from multiple platforms. Moreover, experimental validation underscores HISTEX’s potential to generate new biological insights. It enhances spatial patterns, enriches biologically significant pathways, and facilitates the SR annotation of tissue structures. These findings highlight HISTEX as a powerful tool for advancing ST research. Our source code is available at: <https://github.com/wenwenmin/HISTEX>.

Keywords: Spatial Transcriptomics · Histology Image · Super Resolution · Bidirectional Cross-Attention · Multiple Instance Learning.

1 Introduction

Spatial transcriptomics (ST) has emerged as a groundbreaking technology that enables comprehensive in situ analysis of gene expression profiles within intact tissue architectures, offering unprecedented capabilities for investigating cellular interactions and spatial heterogeneity [22]. While single-cell sequencing is valuable for analyzing the immune cell heterogeneity of disease progression and immune responses, it lacks spatial context, presenting significant limitations in understanding cell-cell interactions and tissue structure [27]. The unique advantages of ST technology have propelled its successful application across diverse

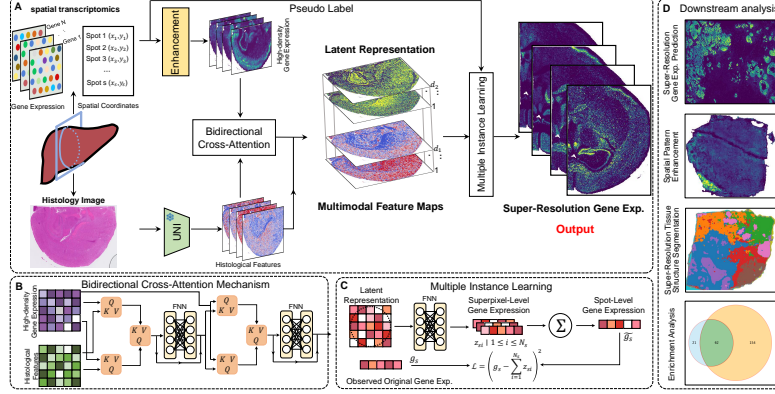


Fig. 1. (A) The overview of HISTEX. (B) The network architecture of bidirectional cross-attention for multimodal fusion. (C) A multi-instance learning framework for model optimization. (D) Biological insights discovered by HISTEX.

biological disciplines, including neuroscience [15], developmental biology [4], and infection and immunity research [24].

Nevertheless, current mainstream ST technologies struggle to balance the spatial resolution of gene expression and the sequencing depth required for comprehensive transcriptome profiling [25]. Imaging-based methodologies, while capable of achieving single-cell or subcellular resolution in gene expression localization, are inherently limited in their transcriptomic scope, typically detecting only a restricted panel of tens to hundreds of target genes [17]. Conversely, NGS-based platforms offer whole-transcriptome analysis and demonstrate superior scalability for large-scale investigations, yet their spatial resolution is constrained by barcode array density limitations [20]. Due to the excessively large spot size and the gaps between spots, a significant amount of gene information is lost across tissue regions, with the observed gene expression representing the aggregated signals from several or even dozens of cells [9,19]. The limited resolution hampers the ability to resolve fine-grained cellular interactions and spatial heterogeneity, diminishing the reliability of ST data for critical applications such as tissue microenvironment analysis and cellular behavior studies.

Several methods have been developed to address the low-resolution (LR) imperfection of NGS-based approaches. STAGE [10] uses a supervised auto-encoder to model the continuity of gene expression in space and generate high-density profiles, but it can only predict the unmeasured gap regions between spots and cannot achieve finer-grained super-resolution generation. Other SR methods [2,28,7,26] achieve the prediction of more fine-grained SR profiles, but they rely solely on histological features without integrating histology and LR spot-based ST data at the input level. As a result, the outcomes of these methods are heavily influenced by morphological similarities across different regions in histological images, leading to deviations from the ground truth.

To address this issue, we develop HISTEX, a multimodal fusion method that accurately generates super-resolution (SR) gene expression profiles by deeply integrating histology images and LR ST data. **The main contributions are summarized as follows:** First, we introduce linear interpolation and a pre-trained foundation model to extract high-density gene expression and informative histological features, maximizing the use of available data. Second, to deeply integrate gene expression and histological features, we introduce the bidirectional cross-attention (BCA), which adaptively aggregates them from multiple perspectives to obtain comprehensive multimodal feature maps, avoiding the limitations of other algorithms that rely on a single data source. Third, Due to the lack of SR-level labels, with each spot containing dozens of SR pixels, we introduce the concept of multi-instance learning (MIL) to optimize the model.

2 Methods

The proposed HISTEX operates in three phases (Fig. 1): (1) **Enhancement and UNI**: extraction of high-density gene expression and histological features through linear interpolation and pre-trained foundation model. (2) **BCA mechanism**: fusion of multimodal representations through a BCA mechanism, enabling deep interaction between transcriptomic and histological domains. (3) **MIL mechanism**: robust generation of SR gene profiles through a MIL framework. Through the aforementioned core framework, HISTEX can effectively integrate histology and spot-based ST data to generate SR gene expression profiles.

2.1 Multimodal Information Enhancement and Extraction

Gene Expression Data Enhancement. The large gaps between spots in ST data lead to substantial gene information loss, disrupting the continuity and statistical significance of gene expression patterns. Therefore, the existing transcriptomic signals were first employed as prior knowledge for predicting high-density gene expression profiles. Let M_g of shape (h, w) be the g -th gene expression matrix, linear interpolation is applied to generate a high-resolution profile:

$$M'_g(i, g) = \frac{1}{2} \left(M_g(i, \lfloor j/2 \rfloor) + M_g(i, \lceil j/2 \rceil) \right), \quad (1)$$

$$M''_g(i, j) = \frac{1}{2} \left(M'_g(\lfloor i/2 \rfloor, j) + M'_g(\lceil i/2 \rceil, j) \right), \quad (2)$$

where M'_g and M''_g have shapes $(2h, w)$ and $(2h, 2w)$. In addition, we introduce the binary mask matrix B , where $B_{i,j} = 1$ denotes a valid spot and $B_{i,j} = 0$ a non-spot region. Then final M_g^{hr} can be obtained:

$$M_g^{\text{hr}} = M''_g \odot B. \quad (3)$$

Histological Feature Extraction. Given the outstanding performance of large pre-trained general-purpose foundation models in clinical tasks [3,12], we

use UNI [3] as the backbone feature extractor. Firstly, the size of histology images were standardized such that each pixel corresponds to $0.5 \mu m$, and the height and width of the entire image are divisible by 224 to accommodate the input requirements of UNI. Let $H = \{H_{ij} \mid H_{ij} \in \mathbb{R}^{224 \times 224}\}$ be the entire histology image, where $i = 1, 2, \dots, \frac{H}{224}$ and $j = 1, 2, \dots, \frac{W}{224}$. UNI partitions each sample into non-overlapping patches of size 16×16 , where each patch is mapped to a 1024-dimensional feature vector. Then the entire histological feature map Y is extracted after each sub-image $H_{i,j}$ is fed into UNI as an input sample:

$$Y_{ij} = \text{UNI}(H_{ij}) \in \mathbb{R}^{14 \times 14 \times 1024}, \quad (4)$$

$$Y = [Y_{ij}]_{i=1, j=1}^{H/224, W/224} \in \mathbb{R}^{M/16 \times N/16 \times 1024}, \quad (5)$$

where Y is formed by concatenating all Y_{ij} according to their original physical positions. We treat each 1024-dimensional vector in Y as a superpixel.

2.2 BCA mechanism for Multimodal Data Fusion

The architecture of cross-attention has shown strong performance in other multimodal tasks [21, 11]. We propose a novel idea for the deep integration of gene expression M^{hr} and histological features Y (refer to Fig. 1B). We illustrate the entire process of multimodal fusion by introducing the basic unit, the Cross-Attention (CA) block, and the BCA layer. We achieve multimodal fusion by stacking two BCA layers, which consists of two parallel CA blocks and one subsequent CA block, where the outputs of the former serve as the output of the latter.

CA Block. The framework of CA block is similar to Multi-head Self-Attention mechanism [23]. The CA block consists of m heads, with the queries Q , keys K , and values V for the i -th head computed as linear transformations of the input representations as follows:

$$Q_i = QW_i^Q, K_i = KW_i^K, V_i = VW_i^V, \quad (6)$$

where the first dimensions of W^Q , W^K , and W^V are determined by the different input data, and the second dimensions represent the embedding length d , which defaults to 512. The operation of the CA block can be described as follows:

$$\begin{aligned} h_i &= A(Q_i, K_i, V_i) = \text{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_h}}\right) V_i, \\ h &= h_1 \oplus h_2 \oplus \dots \oplus h_m, \\ \text{CA}(Q, K, V) &= hW, \end{aligned} \quad (7)$$

where \oplus represents concatenation of heads.

BCA Layer. Let $H_j \in \mathbb{R}^{n_1, 1024}$ denote the histological feature Y , where n_1 represents the number of superpixel spanned by spot, $G_j \in \mathbb{R}^{n_2, g}$ denote the gene expression data obtained from M^{hr} , where g represents the number of

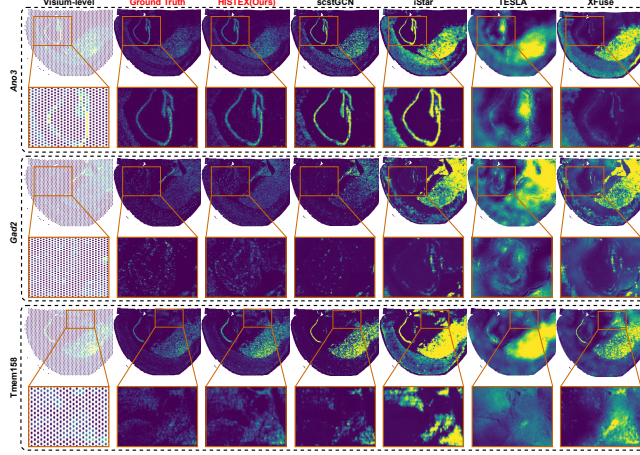


Fig. 2. Visualization of several disease-related genes with different spatial patterns in MBHD data for comparison between HISTEX and other state-of-the-art (SOTA) methods.

genes, and G_j is selected based on the Euclidean distance to the closest H_j . The process of three CA blocks working collaboratively is as follows:

$$\begin{aligned} Z_{j1} &= \text{CA}(G_j, H_j, H_j), \\ Z_{j2} &= \text{CA}(H_j, G_j, G_j), \\ Z_{j3} &= \text{CA}(Z_{j2}, Z_{j1}, Z_{j1}). \end{aligned} \quad (8)$$

Next, the output of the final CA block is passed through the FNN and a residual connection [6], forming the output of the BCA layer:

$$\text{BCA}(G_j, H_j) = \text{FFN}(Z_{j3}) + Z_{j3}. \quad (9)$$

HISTEX consists of two connected BCA layers, where the output of the first layer replaces H_j and serves as part of the input to the second layer. After the above process, stacking with H_j results in the multimodal feature map:

$$L_j = \text{BCA}(G_j, \text{BCA}(G_j, H_j)) + H_j, \quad (10)$$

where L_j has a shape of $(n_1, 1024+d)$, and the final L is formed by reconstructing L_j into a 3D structure with shape $(M/16, N/16, 1024+d)$.

2.3 Model Optimization by MIL

HISTEX is trained using the concept of MIL, where each spot represents a bag, and the superpixels covered by the spot are treated as instances (Fig. 1C). The instance-level gene expression are then aggregated to obtain the bag-level gene

Table 1. Numerical evaluation experiments comparing HISTEX with baselines.

Methods	MBHD		HBCHD		HBC_1		HBC_2		HBC_3	
	RMSE	SSIM	RMSE	SSIM	RMSE	SSIM	RMSE	SSIM	RMSE	SSIM
STAGE [10]	0.1638	0.3561	0.1950	0.3762	0.1483	0.4565	0.2130	0.4428	0.1648	0.3965
XFuse [2]	0.1236	0.3894	0.1439	0.4268	0.1836	0.4114	0.1495	0.3834	0.1536	0.3452
TESLA [7]	0.1004	0.6136	0.1264	0.6237	0.0984	0.5138	0.0862	0.5483	0.0943	0.6476
iStar [28]	0.0797	0.6934	0.1183	0.6976	0.0817	0.5851	0.0672	0.6839	0.0746	0.6298
sctGCN [26]	0.0542	0.7409	0.0617	0.7462	0.0601	0.7134	0.0591	0.7461	0.0653	0.6834
Ours	0.0292	0.8443	0.0352	0.8273	0.0365	0.8234	0.0349	0.8493	0.0382	0.7957

expression. Let S be the number of spots in M^{hr} , $y_s \in \mathbb{R}^{1,g}$ be the gene expression at spot s , $L_s \in \mathbb{R}^{n,1024+d}$ be the set of superpixels covered by spot s , $e_c \in \mathbb{R}^{1,n}$ be the unit vector used for aggregation operations. Then the loss function is:

$$\mathcal{L} = \sum_{s=1}^S \left(y_s - e_c \text{FFN}(L_s) \right)^2. \quad (11)$$

3 Experimental Results

3.1 Dataset and Implementation

Dataset Preprocessing. Since the spot-based ST data lacks SR gene expression as ground truth, we conducted a comparative numerical evaluation of HISTEX and other baselines on multiple data from the Visium HD and Xenium platforms. For the Visium HD data, we performed aggregation on the raw data (with bins of $8 \times 8 \mu m^2$) based on the distribution patterns and sizes of the spots in the Visium data, in order to generate pseudo-Visium as inputs. For the Xenium data, we first constructed a rectangular grid with each cell measuring $8 \times 8 \mu m^2$. Based on the overlap area between each grid cell and the cells in the Xenium data, we reshaped the raw data into a regular gene expression as the ground truth. Next we generated pseudo-Visium data following a procedure similar to that for the Visium HD. In our experiments, we predicted the top 1,000 highly variable genes (HVGs) in each Visium HD dataset and all genes in the Xenium dataset (313, 288, and 377 genes for the three sections, respectively).

Dataset Source. The Xenium human breast cancer (HBC) datasets with three sections (denoted as HBC_1, HBC_2, and HBC_3) can be accessed at <https://www.10xgenomics.com/products/xenium-in-situ/preview-dataset-human-breast> [8]. The Visium HD human breast cancer (HBCHD) and mouse brain (MBHD) are available at <https://www.10xgenomics.com/datasets/visium-hd-cytassist-gene-expression-human-breast-cancer-fresh-frozen> and <https://www.10xgenomics.com/datasets/visium-hd-cytassist-gene-expression-mouse-brain-fresh-frozen> [13], respectively. The HER2-positive breast cancer (HER2ST) datasets can be found at <https://github.com/almaan/her2st>.

Implementation Details. All experiments were conducted on an NVIDIA RTX 3090 GPU, using the Pytorch 2.1.1 and Python 3.11.5 environment, with a total of 500 training epochs and a learning rate of 0.0001, employing the L_1

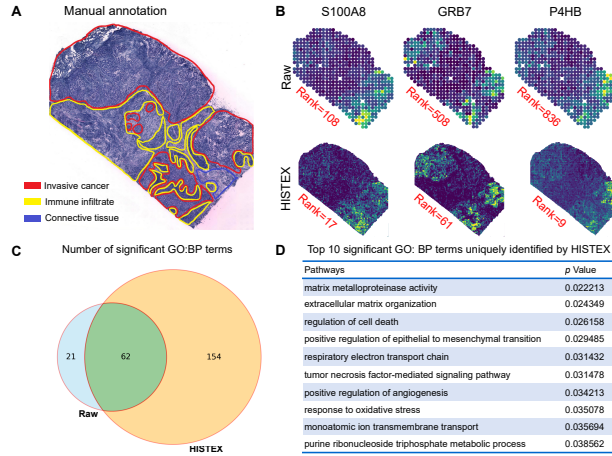


Fig. 3. (A) Manual annotation of section E1 in HER2ST. (B) Spatial pattern enhancement. (C) Enrichment analysis. (D) Top 10 significant terms are found in HISTEX.

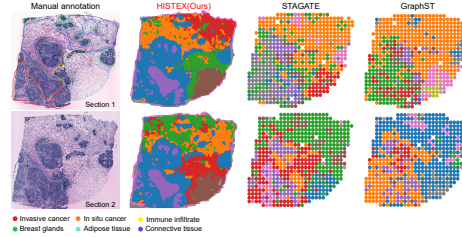


Fig. 4. The outstanding SR tissue annotation capability of HISTEX.

loss and the Adam optimizer. For bidirectional cross-attention mechanism, we performed deep information fusion by selecting 6 spots for each superpixel, based on their Euclidean distance metric.

3.2 Results

Evaluation of Generation Performance. Numerical evaluation experiments were conducted on two Visium HD datasets (HBCHD and MBHD), and a Xenium dataset (HBC). We used pseudo-Visium data and histology images as the input to HISTEX and conducted a comprehensive comparison of its SR gene expression prediction performance with other baselines, evaluated using the root mean square error (RMSE) and structural similarity index measure (SSIM) metrics (refer to Section 3.1). Before computing the evaluation metrics, we normalized the intensities of both the ground truth and the predicted super-resolution gene expression maps to the range of $[0, 1]$. The results show that HISTEX achieved the predictions closest to the ground truth across all datasets, outper-

Table 2. Comparison of different model variants in ablation study.

Methods	MBHD		HBCHD		HBC_1		HBC_2		HBC_3	
	RMSE	SSIM	RMSE	SSIM	RMSE	SSIM	RMSE	SSIM	RMSE	SSIM
w/o His-Fea	0.0836	0.6446	0.1368	0.6534	0.0946	0.5639	0.0835	0.6793	0.0872	0.6016
w/o BCA	0.0586	0.7291	0.0628	0.7226	0.0634	0.6794	0.0637	0.7139	0.0715	0.6792
w/o Ge-En	0.0397	0.8255	0.0484	0.7856	0.0513	0.6667	0.0495	0.0713	0.0548	0.6758
Ours	0.0292	0.8443	0.0352	0.8273	0.0365	0.8234	0.0349	0.8493	0.0382	0.7957

forming all baselines (Table 1). Compared to SOTA method, HISTEX shows improvements ranging from 39.2% to 46.1% in RMSE and 10.8% to 16.4% in SSIM across different datasets. Next, we selected several disease-related genes with distinct spatial patterns for visualization to compare the performance of HISTEX with SOTA methods (Fig. 2). Although all methods can infer the SR gene expression, the baselines exhibit a significant deviation from the ground truth. Both globally and locally, HISTEX captures the true spatial patterns more accurately, with the distribution of numerical values being closer to the ground truth.

Insights from Downstream Analysis. We analyzed the capabilities of HISTEX in discovering new biological insights using HER2ST dataset (Fig. 3A). The original data, limited by low resolution, disrupts the continuity of spatial patterns. After enhancement with HISTEX, following the Sepal-based criterion for ranking gene spatial patterns [1], the spatial patterns of certain disease-related genes are significantly increased, exhibiting more statistically meaningful patterns (Fig. 3B). Additionally, 154 GO:BP terms were detected in the SR data generated by HISTEX but not enriched in the original data [18] (Fig. 3C). Furthermore, among the top 10 significant terms, several are related to malignancies [16,14], providing new insights into the identification of disease mechanisms (Fig. 3D). Finally, we evaluated the ability of HISTEX in spatial domain identification. Traditional spatial clustering methods are limited to tissue annotation in LR scenarios. In contrast, after improving the resolution of gene expression, HISTEX leverages the K-means [5] to perform tissue annotation under SR conditions. The results show that HISTEX effectively identifies regions such as Breast glands and In situ cancer, outperforming other methods (Fig. 4).

Ablation Study. We evaluated the contribution of several key components in HISTEX. 1) w/o His-Fea: Relying solely on the output of the BCA module for prediction, without utilizing histological features. 2) w/o BCA: Completely removing the multimodal fusion module (BCA) and relying solely on histological features for prediction. 3) w/o Ge-En: Removing the Gene Expression Enhancement module and directly using the original LR data as input to the BCA module. The results show that the performance of all modules is worse than HISTEX in multiple datasets, which means that these modules have improved the performance of HISTEX. The results show that the performance of all models is inferior to HISTEX across multiple datasets, indicating that each of these modules contributes to enhancing the performance of HISTEX.

4 Conclusion

In this study, we present HISTEX, a novel multimodal information fusion model designed to predict SR spatial gene expression. The first step of HISTEX is to generate high-density gene expression and rich histological features. The second step involves deep multimodal information fusion through BCA mechanism proposed in this paper. Finally, SR spatial gene expression profiles are predicted by HISTEX trained with a MIL framework. Numerical evaluation experiments and spatial visualization results on multiple datasets demonstrate that HISTEX outperforms other SOTA methods. Moreover, HISTEX can also offer new insights in biomedical research, such as gene expression pattern enhancement, enrichment of biologically significant pathways, and SR annotation of tissue structure, facilitating a deeper understanding of biological processes for researchers.

Acknowledgments. The work was supported in part by the National Natural Science Foundation of China (62262069), Yunnan Fundamental Research Projects (202301AT070230) and Young Talent Program of Yunnan Province (C619300A067).

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

References

1. Andersson, A., Lundeberg, J.: Sepal: Identifying transcript profiles with spatial patterns by diffusion-based modeling. *Bioinformatics* **37**(17), 2644–2650 (2021)
2. Bergenstr hle, L., He, B., Bothers: Super-resolved spatial transcriptomics by deep data fusion. *Nature Biotechnology* **40**(4), 476–479 (2022)
3. Chen, R.J., Ding, T., et al.: Towards a general-purpose foundation model for computational pathology. *Nature Medicine* **30**(3), 850–862 (2024)
4. Choe, K., Pak, U., et al.: Advances and challenges in spatial transcriptomics for developmental biology. *Biomolecules* **13**(1), 156 (2023)
5. Hamerly, G., Elkan, C.: Learning the k in k-means. *Advances in Neural Information Processing Systems* **16**, 281–288 (2003)
6. He, K., Zhang, X., et al.: Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 770–778 (2016)
7. Hu, J., Coleman, K., et al.: Deciphering tumor ecosystems at super resolution from spatial transcriptomics with tesla. *Cell Systems* **14**(5), 404–417 (2023)
8. Janesick, A., Shelansky, R., et al.: High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. *Nature Communications* **14**(1), 8353 (2023)
9. Ji, A.L., Rubin, A.J., et al.: Multimodal analysis of composition and spatial architecture in human squamous cell carcinoma. *Cell* **182**(2), 497–514 (2020)
10. Li, S., Gai, K., et al.: High-density generation of spatial transcriptomics with stage. *Nucleic Acids Research* **52**(9), 4843–4856 (2024)

11. Li, Y., Yu, A.W., et al.: Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17182–17191 (2022)
12. Lu, M.Y., Chen, B., et al.: A visual-language foundation model for computational pathology. *Nature Medicine* **30**, 863–874 (2024)
13. Nagendran, M., Sapida, J., et al.: 1457 visium HD enables spatially resolved, single-cell scale resolution mapping of ffpe human breast cancer tissue (2023)
14. Niland, S., Riscanevo, A.X., Eble, J.A.: Matrix metalloproteinases shape the tumor microenvironment in cancer progression. *International journal of molecular sciences* **23**(1), 146 (2021)
15. Park, H.E., Jo, S.H., et al.: Spatial transcriptomics: technical aspects of recent developments and their applications in neuroscience and cancer research. *Advanced Science* **10**(16), 2206939 (2023)
16. Peng, F., Liao, M., et al.: Regulated cell death (rcd) in cancer: key pathways and targeted therapies. *Signal transduction and targeted therapy* **7**(1), 286 (2022)
17. Petukhov, V., Xu, R.J., et al.: Cell segmentation in imaging-based spatial transcriptomics. *Nature Biotechnology* **40**(3), 345–354 (2022)
18. Raudvere, U., Kolberg, L., et al.: g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research* **47**(W1), W191–W198 (2019)
19. Shi, Z., Xue, S., et al.: High-resolution spatial transcriptomics from histology images using histosge. In: 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 2402–2407. IEEE (2024)
20. Si, Y., Lee, C., et al.: Ficture: scalable segmentation-free analysis of submicron-resolution spatial transcriptomics. *Nature Methods* **21**(3), 1843–1854 (2024)
21. Tan, H., Bansal, M.: Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490* (2019)
22. Tian, L., Chen, F., Macosko, E.Z.: The expanding vistas of spatial transcriptomics. *Nature Biotechnology* **41**(6), 773–782 (2023)
23. Vaswani, A., Shazeer, N., et al.: Attention is all you need. *Advances in Neural Information Processing Systems* **14**, 1–11 (2017)
24. Williams, C.G., Lee, H.J., et al.: An introduction to spatial transcriptomics for biomedical research. *Genome Medicine* **14**(1), 68 (2022)
25. Xue, S., Zhu, F., et al.: Stentrans: transformer-based deep learning for spatial transcriptomics enhancement. In: International Symposium on Bioinformatics Research and Applications. pp. 63–75. Springer (2024)
26. Xue, S., Zhu, F., et al.: Inferring single-cell resolution spatial gene expression via fusing spot-based spatial transcriptomics, location, and histology using gcn. *Briefings in Bioinformatics* **26**(1), bbae630 (2025)
27. You, Y., Fu, Y., et al.: Systematic comparison of sequencing-based spatial transcriptomic methods. *Nature Methods* **21**(9), 1743–1754 (2024)
28. Zhang, D., Schroeder, A., et al.: Inferring super-resolution tissue architecture by integrating spatial transcriptomics with histology. *Nature Biotechnology* pp. 1–6 (2024)