# C3-Diff: Super-resolving Spatial Transcriptomics via Cross-modal Cross-content Contrastive Diffusion Modelling

Xiaofei Wang
University of Cambridge
xw405@cam.ac.uk

Stephen J Price
University of Cambridge
sjp58@cam.ac.uk

Chao Li
University of Cambridge
University of Dundee
cl647@cam.ac.uk

## 1. Mutual Information Maximization Analysis

### 1.1. Content-invariant Features by $\mathcal{L}_{\text{content}}$

Here we demonstrate that the proposed $\mathcal{L}_{\text{modal}}$ can help the model to learn content-invariant features, via the analysis of mutual information maximization. Specifically, mutual information captures the nonlinear statistical dependencies between variables. For cross-content contrastive loss $\mathcal{L}_{\text{content}}$, the mutual information for the positive pair $(z, z^+)$ is defined as

$$\begin{aligned} I(z, z^+) &= \sum_{z, z^+} p(z, z^+) \log \frac{p(z, z^+)}{p(z)p(z^+)} \\ &= \sum_{z, z^+} p(z, z^+) \log \frac{p(z|z^+)}{p(z)}. \end{aligned}$$

In the above equation, $(z, z^+) \sim ([\mathbf{C}_{\text{h}}]_j, [\hat{\mathbf{C}}_{\text{y}}]_j)$, so this equation can be re-write as

$$\begin{aligned} &I(z, z^+)_{\text{content}} \\ &= \sum_{[\mathbf{C}_{\text{h}}]_j, [\hat{\mathbf{C}}_{\text{y}}]_j} p([\mathbf{C}_{\text{h}}]_j, [\hat{\mathbf{C}}_{\text{y}}]_j) \log \frac{p([\mathbf{C}_{\text{h}}]_j, [\hat{\mathbf{C}}_{\text{y}}]_j)}{p([\mathbf{C}_{\text{h}}]_j)p([\hat{\mathbf{C}}_{\text{y}}]_j)} \\ &= \sum_{[\mathbf{C}_{\text{h}}]_j, [\hat{\mathbf{C}}_{\text{y}}]_j} p([\mathbf{C}_{\text{h}}]_j, [\hat{\mathbf{C}}_{\text{y}}]_j) \log \frac{p([\mathbf{C}_{\text{h}}]_j|[\hat{\mathbf{C}}_{\text{y}}]_j)}{p([\mathbf{C}_{\text{h}}]_j)}. \end{aligned}$$

where $p([\mathbf{C}_{\text{h}}]_j|[\hat{\mathbf{C}}_{\text{y}}]_j)/p([\mathbf{C}_{\text{h}}]_j)$ represents the density ratio between the two elements of the positive pair. According to the proof in [6, 7], the optimization of contrastive loss based on maximum likelihood estimation is equal to estimating the density ratio in positive training pairs. Therefore, with the minimization of $\mathcal{L}_{\text{content}}$, we can achieve mutual information maximization between the positive pair of $(z, z^+)$. Besides, the original images of the positive pair $[\mathbf{C}_{\text{h}}]_j$ and $[\hat{\mathbf{C}}_{\text{y}}]_j$ are from the same tissue content yet with different modalities of histology image and ST maps. Therefore, the proposed $\mathcal{L}_{\text{content}}$ can force the model to learn tissue content-invariant features for ST enhancement.

### 1.2. Joint Content- & Modality-invariant Features by L_inter-sphere

In addition, the proposed $\mathcal{L}_{\text{inter-sphere}}$ can also help the model to learn both content-invariant and modal-invariant features. Similarly, for inter-sphere contrastive loss $\mathcal{L}_{\text{inter-sphere}}$, the mutual information for the positive pair $(z, z^+) \sim ([\mathbf{M}_{\text{h}}]_j, [\mathbf{C}_{\text{h}}]_j)$ is defined as

$$\begin{aligned} &I(z, z^+)_{\text{inter-sphere}} \\ &= \sum_{[\mathbf{M}_{\text{h}}]_j, [\mathbf{C}_{\text{h}}]_j} p([\mathbf{M}_{\text{h}}]_j, [\mathbf{C}_{\text{h}}]_j) \log \frac{p([\mathbf{M}_{\text{h}}]_j, [\mathbf{C}_{\text{h}}]_j)}{p([\mathbf{M}_{\text{h}}]_j)p([\mathbf{C}_{\text{h}}]_j)} \\ &= \sum_{[\mathbf{M}_{\text{h}}]_j, [\mathbf{C}_{\text{h}}]_j} p([\mathbf{M}_{\text{h}}]_j, [\mathbf{C}_{\text{h}}]_j) \log \frac{p([\mathbf{M}_{\text{h}}]_j|[\mathbf{C}_{\text{h}}]_j)}{p([\mathbf{M}_{\text{h}}]_j)}. \end{aligned}$$

where $p([\mathbf{M}_{\text{h}}]_j|[\mathbf{C}_{\text{h}}]_j)/p([\mathbf{M}_{\text{h}}]_j)$ represents the density ratio between the two elements of the positive pair. Therefore, with the minimization of $\mathcal{L}_{\text{inter-sphere}}$, we can achieve mutual information maximization between the positive pair of $(z, z^+)$. Of note, the original images of the positive pair $[\mathbf{M}_{\text{h}}]_j$ and $[\mathbf{C}_{\text{y}}]_j$ are from the same tissue content and modality, but are derived from different encoders. Hence, the proposed $\mathcal{L}_{\text{inter-sphere}}$ can force the model to learn both tissue content-invariant and modal-invariant features for ST enhancement.

## 2. Overall Loss

By combining the basic loss $\mathcal{L}_{\text{mse}}$ of the diffusion model and the cross-modal cross-content contrastive loss, the overall loss function for our C3-Diff framework can be formulated as

Table 1. Implementation details of our proposed method.

| | |
|---|---|
| Number of genes included in this study | 200 |
| Initial value of imputation adjusting factor $\alpha$ | 1 |
| Initial value of imputation adjusting factor $\beta$ | 1 |
| $\lambda_{\text{mse}}$ for $\mathcal{L}_{\text{mse}}$ | 2 |
| $\lambda_{\text{modal}}$ for $\mathcal{L}_{\text{modal}}$ | 1 |
| $\lambda_{\text{content}}$ for $\mathcal{L}_{\text{content}}$ | 1 |
| $\lambda_{\text{inter-sphere}}$ for $\mathcal{L}_{\text{inter-sphere}}$ | 1 |
| Exponential decay rate $\beta_1$ and $\beta_2$ for AdamW optimization | 0.9 and 0.999 |
| Epsilon $\epsilon$ for AdamW optimization | $1 \times 10^{-8}$ |
| Weight decay for AdamW optimization | $1 \times 10^{-5}$ |

$$\mathcal{L}_{\text{overall}} = \lambda_{\text{mse}}\mathcal{L}_{\text{mse}} + \lambda_{\text{modal}}\mathcal{L}_{\text{modal}} + \lambda_{\text{content}}\mathcal{L}_{\text{content}} + \lambda_{\text{inter-sphere}}\mathcal{L}_{\text{inter-sphere}}$$

where $\lambda_{\text{mse}}, \lambda_{\text{modal}}, \lambda_{\text{content}}$ and $\lambda_{\text{inter-sphere}}$ are hyper-parameters to balance the diffusion loss, cross-modal contrastive loss, cross-content contrastive loss and inter-sphere contrastive Loss. Of note, the modality incomplete samples only contribute to the optimization of our basic loss $\mathcal{L}_{\text{mse}}$ of the diffusion model.

## 3. Datasets and Implementation Details

### 3.1. Dataset Preparation

Overall, we evaluate our model on four public datasets, i.e., Breast-Xenium [3], Melanoma-Xenium [8], Breast-SGE [4], and Breast-ST [1]. In all datasets, we use LR ST maps and paired HR histology images to restore $5\times$ and $10\times$ HR ST maps, consistent with settings in [2, 9]. For downsampling, we use standard bicubic interpolation. In addition, with the gene panels (280 genes) in the four datasets, we choose the top 200 highly variable genes for our task. Totally, we include 514 histology images and 102,800 ST maps across 200 genes. See below for more details of dataset preparation and evaluation metrics.

All train/val/test splits are performed at the patient level across all datasets, so no patient appears in multiple splits (i.e., no leakage). For HVG selection, we follow standard practice by selecting the top-$N$ HVGs (as in widely used pipelines such as Seurat and Scanpy) to reduce low-variance noise. For the 280-gene setting, we use $N = 200$ (70%) as a simple, round cutoff.

**Breast-Xenium Dataset:** We randomly split the 232 histology images of human breast cancer (with 46,400 ST maps) into 99 (with 19,800 ST maps) for training, 49 (with 9,800 ST maps) for validation, and 84 (with 16,800 ST maps) for test. For both SR times, the histology images are of 256 $\times$ 256 pixels at 10 $\mu$m px$^{-1}$, while the LR ST maps are of 26 $\times$ 26 pixels at 100 $\mu$m px$^{-1}$. Besides, the HR ST maps are of 256 $\times$ 256 pixels at 10 $\mu$m px$^{-1}$ and of 128 $\times$ 128 pixels at 20 $\mu$m px$^{-1}$ for $10\times$ and $5\times$ SR, respectively.

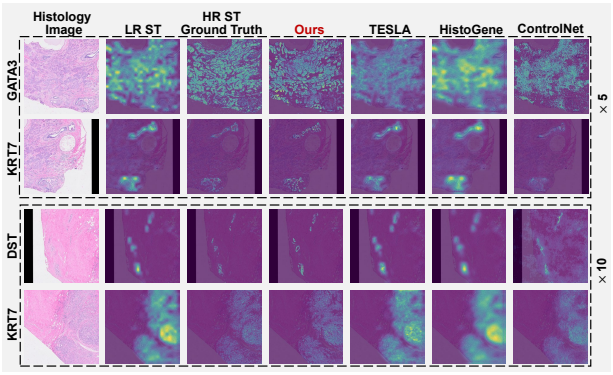


Figure 1. Additional visual comparisons at $5\times$ and $10\times$ scales on the Breast-Xenium dataset. The ST maps are overlayed on the paired histology image for better visualisation. Note that KRT7, GATA3 and DST denote different genes.

**Melanoma-Xenium Dataset:** We randomly split the 12 histology images of melanoma (with 2,400 ST maps) into 7 (with 1,400 ST maps) for training, 2 (with 400 ST maps) for validation, and 3 (with 600 ST maps) for test. Besides, the resolution setting is the same as Breast-Xenium.

**Two External Validation Datasets:** The Breast-SGE (47 histology and 9,400 ST) and Breast-ST (223 histology and 44,600 ST) are with the same resolution setting as Breast-Xenium and are both set as external validation datasets.

**Evaluation Metrics:** We use Root MSE (RMSE) and Pearson correlation coefficient (PCC) to evaluate the performance of ST enhancement. For Breast-Xenium and Melanoma-Xenium datasets, we have HR ST as ground truth, while the HR ST is unavailable in two external validation sets, so we follow [2] for model evaluation, where ST enhancement should retain the original spot level pattern while increasing resolution. Specifically, the enhanced ST maps are first downsampled to LR, and then used to calculate metrics with paired LR ST.

### 3.2. Implementation Details

We use ControlNet as our backbone, with *off-the-shelf representation* of the unconditional image synthesis task on CelebA-HQ dataset [11]. We train our model for 50 epochs on two NVIDIA RTX A5000 24 GB GPUs, with batch size 16 and learning rate 0.00001 with AdamW optimizer [5]. Following [10], C3-Diff uses the sampling steps of 50 and eta of 0. See other key hyper-parameters in the Table 1. The contrastive learning and ControlNet are trained jointly. All hyper-parameters are tuned to achieve the best performance over the validation set. Our method is implemented on PyTorch with the Python environment.

## 4. Cell Type Localization

Assume that the total number of genes incorporated in the model is $K$ ($K$=200 in our paper). Let $T$ ($T$=7 in our paper) be the total number of candidate cell types, which include 7 type of immune-related cells of B cell,T cell CD4,T Cell CD8, T cells regulatory (Tregs), Macrophages M0 cell, Macrophages M2 cell and plasma cell. For each cell type $t \in \{1, ..., T\}$, suppose we have a list of marker gene indices $I - t$, which is a subset of $\{1, ..., K\}$. For each marker gene $k \in I_t$, we standardize its predicted super-resolution gene expression image $X_k \in \mathbb{R}^{256\times256}$ into the range of $[0.0, 1.0]$ and obtain $\hat{X}_k \in \mathbb{R}^{256\times256}$, where $\hat{X}_k = (X_k - \min X_k)/(\max X_k - \min X_k)$. Then for each pixel $(i, j)$ in $\hat{X}_k$, we compute the score for cell type $t$ by averaging the standardized gene expressions of all its marker genes:

$$s_t^{i,j} = |I_t|^{-1} \sum_{k\in I_t} \hat{X}_k^{i,j}, \quad \text{where} |I_t| \text{is the number of genes in} I_t \tag{1}$$

Finally, the cell type with the maximal probability score can be obtained as $t_{max}^{i,j} = \text{argmax}_{1\le t\le T} s_t^{i,j}$, with $s_{max}^{i,j} = \max_{1\le t\le T} s_t^{i,j}$ as the score of this cell type.

## References


[1] Bryan He, Ludvig Bergenstråhle, Linnea Stenbeck, Abubakar Abid, Alma Andersson, Åke Borg, Jonas Maaskola, Joakim Lundeberg, and James Zou. Integrating spatial gene expression and breast tumour morphology via deep learning. *Nature biomedical engineering*, 4(8): 827–834, 2020. 2

[2] Jian Hu, Kyle Coleman, Daiwei Zhang, Edward B Lee, Humam Kadara, Linghua Wang, and Mingyao Li. Deciphering tumor ecosystems at super resolution from spatial transcriptomics with tesla. *Cell systems*, 14(5):404–417, 2023. 2

[3] Amanda Janesick, Robert Shelansky, Andrew D Gottscho, Florian Wagner, Stephen R Williams, Morgane Rouault, Ghezal Beliakoff, Carolyn A Morrison, Michelli F Oliveira, Jordan T Sicherman, et al. High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. *Nature Communications*, 14(1):8353, 2023. 2

[4] Tapsi Kumar, Kevin Nee, Runmin Wei, Siyuan He, Quy H Nguyen, Shanshan Bai, Kerrigan Blake, Maren Pein, Yanwen Gong, Emi Sei, et al. A spatially resolved single-cell genomic atlas of the adult human breast. *Nature*, 620(7972): 181–191, 2023. 2

[5] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 3

[6] Hiroaki Sasaki and Takashi Takenouchi. Representation learning for maximization of mi, nonlinear ica and nonlinear subspaces with robust density ratio estimation. *Journal of Machine Learning Research*, 23(231):1–55, 2022. 1

[7] Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density ratio estimation in machine learning*. Cambridge University Press, 2012. 1

[8] Huan Wang, Ruixu Huang, Jack Nelson, Ce Gao, Miles Tran, Anna Yeaton, Kristen Felt, Kathleen L Pfaff, Teri Bowman, Scott J Rodig, et al. Systematic benchmarking of imaging spatial transcriptomics platforms in ffpe tissues. *bioRxiv*, 2023. 2

[9] Daiwei Zhang, Amelia Schroeder, Hanying Yan, Haochen Yang, Jian Hu, Michelle YY Lee, Kyung S Cho, Katalin Susztak, George X Xu, Michael D Feldman, et al. Inferring super-resolution tissue architecture by integrating spatial transcriptomics with histology. *Nature Biotechnology*, pages 1–6, 2024. 2

[10] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 3

[11] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. Celebv-hq: A large-scale video facial attributes dataset. In *European conference on computer vision*, pages 650–667. Springer, 2022. 3