SSL-CST: Cell Segmentation for Single-Cell Spatial Transcriptome Based on Self-Supervised Learning

Weiliang Huo[1], Shilin Zhang[2], Suixue Wang[1], Qingchen Zhang[3, *]
[1]School of Information and Communication Engineering, Hainan University
[2]College of Intelligence and Computing, Tianjin University
[3]School of Computer Science and Technology, Hainan University
{wlhuo, wangsuixue, zhangqingchen}@hainanu.edu.cn, zhang_shilin_sd@163.com

Abstract

The continuous advancements in life science technology have enabled spatial transcriptome technology to achieve an impressive level of resolution at the single-cell level. This technology has emerged as a crucial method for studying the cellular composition and differentiation states of tissues, investigating cell-cell interactions, and unraveling the molecular mechanisms underlying diseases and developmental processes. A key component in this analysis is the accurate segmentation of cells. However, existing segmentation methods often fail to fully leverage the valuable information provided by spatial transcriptomics, leading to inaccurate cell segmentation. In this study, we introduce SSL-CST, a cell segmentation for single-cell spatial transcriptome method based on self-supervised learning. SSL-CST employs a pre-trained model for foundational contour segmentation. Following the denoising process, it utilizes a self-supervised neural network to correct the cell boundaries to obtain accurate cell boundaries. Through this approach, SSL-CST outperforms other state-of-the-art methods in various tests conducted on multiple datasets. The improved segmentation provided by SSL-CST further enhances the analysis of single-cell spatial expression, providing effective tools for biological discovery.

Code — https://github.com/wlhuo/SSL-CST

Introduction

Cancer, a significant challenge of the 21st century, poses substantial social, public health, and economic burdens(Bray et al. 2024). It is responsible for a sizeable amount, about 22.8%, of deaths globally attributable to noncommunicable diseases(Sung et al. 2021). Disturbingly, the incidence of new cancer cases is projected to rise to around 30 million globally by 2040, exerting immense pressure on healthcare systems and communities(Chhikara, Parang et al. 2023). A series of studies have consistently demonstrated that high heterogeneity is a major obstacle in effectively treating tumors(Bhang et al. 2015). Consequently, characterizing tumor heterogeneity is paramount for advancing our knowledge of cancer biology. Indeed, inter-tumor heterogeneity and intra-tumor heterogeneity are the two primary categories into which tumor heterogeneity can be generally divided. Inter-tumor heterogeneity refers to the variations observed between tumors in different patients. A critical necessity in clinical practice is the single-cell identification of cell types and their distribution within the tumor microenvironment(Dagogo-Jack and Shaw 2018). This comprehensive understanding can unravel the mechanisms underlying cancer occurrence and development, and ultimately provide valuable strategies and methods for clinical management. Unraveling cellular heterogeneity has become easier with the use of single-cell sequencing techniques(Gohil et al. 2021). However, it is difficult to characterize intra-spatial cell interactions within tumors because this technique causes the loss of spatial histology information during the cell separation procedure(Tellez-Gabriel et al. 2016). This loss of spatial context hampers the ability to comprehend the intricate cellular relationships and communication networks that exist within the tumor microenvironment.

An increasing number of studies are utilizing spatial transcriptome analysis to investigate the spatial heterogeneity of cancer(Zheng and Fang 2022). Based on the methods of detection used, spatial transcriptome technology can be divided into two main categories: sequencing-based methods(Zhuang 2021) and in situ-based methods(Liu et al. 2020). These approaches cover a range of cellular resolutions, from subcellular to multicellular levels(Asp, Bergenstråhle, and Lundeberg 2020). In the in situ-based method, high-precision spatial transcriptomic sequencing technology(Racle et al. 2017) offers nanoscale resolution for detection. By combining a DNA nanosphere pattern array chip and RNA in situ hybridization, this method can capture hundreds of cells at each spatial point, enabling accurate identification of cell types at the single-cell level and providing insights into the cell distribution and composition within the tumor microenvironment. Methods that convert raw spatial transcriptomics data into spatial expression profiles at the single-cell level are integral to the process of cell type identification. As it allows for precise localization of cell positions and regions, thereby enhancing the overall accuracy of spatial transcriptomic technology. Commonly used methods usually utilize transcriptome data to classify individual transcriptomes for cell segmentation, such as Baysor(Petukhov et al. 2022), Sparcle(Prabhakaran 2022),

*Corresponding authors: Qingchen Zhang

and ClusterMap(He et al. 2021), frequently operate under the assumption that RNA expression is uniform across the entire cell. This foundational assumption represents a significant limitation of these approaches, as it overlooks the inherent spatial and functional heterogeneity that can exist within a single cell. With the continuous development of analysis tools and methods, researchers have recognized the importance of incorporating auxiliary staining methods and histological images into the cell segmentation process(Wang et al. 2019). JSTA(Littman et al. 2021) enhances the segmentation results by integrating cell type probabilities for each pixel and iteratively refining the assignment of boundary pixels based on these probabilities. However, the effectiveness of JSTA diminishes under conditions of low segmentation quality. SCS(Chen, Li, and Bar-Joseph 2023) employs a comprehensive strategy that integrates staining images with spatial transcriptome gene expression data to enhance the accuracy of cell segmentation. However, SCS selects only spots with high transcriptome expression intensity as positive samples during the sample selection process. While this method effectively identifies regions of interest, it may be susceptible to biases introduced by technical artifacts or random fluctuations in transcript abundance. Although existing methods have shown promising results, they may not fully exploit the expression information from spatial transcriptomics data and staining image information to address the cell segmentation problem.

Aiming at this problem, we present SSL-CST, a cell segmentation method for single-cell spatial transcriptome based on self-supervised learning. The main contributions of this work are summarized as follows:

1) We performed initial contour segmentation using a pretrained model, combined with a dynamic adaptive morphological erosion algorithm for denoising. Additionally, we employed a self-supervised network to refine cell boundaries, thereby overcoming the limitations of traditional methods that rely heavily on labeled data and struggle to capture cellular heterogeneity.

2) We innovatively integrated positional information from nuclear staining images, gene expression intensity, and spatial distance features to construct a multi-dimensional input. This was achieved through a hybrid encoding architecture that combines self-attention mechanisms with a reparameterized multilayer perceptron(RMLP) for feature extraction, effectively leveraging the multi-source information inherent in spatial transcriptomics data.

3) Experimental results demonstrate that our method outperforms state-of-the-art approaches across multiple high-resolution spatial transcriptomics datasets(SeqFISH+(Eng et al. 2019), Stereo-seq(Chen et al. 2022), and Merscope-Ovarian(Huang et al. 2024)). This advancement provides a more reliable tool for analyzing the tumor microenvironment.

Material and Methods

The data required for the method is segregated into two components: spatial transcriptomics sequencing data and corresponding nuclear staining images. The sequencing data should encompass essential information, including the gene type, coordinates, and gene expression intensity.

Nuclei segmentation

The absence of ground truth labels for cells in spatial transcriptomic staining images and the different imaging methods of different source data renders the training of models or the fine-tuning of pre-trained models exceedingly virtually impossible. We employed pre-trained models for nuclei segmentation and assessed the performance of four prominent cell segmentation frameworks: Cellpose(Pachitariu and Stringer 2022), DeepCell(Bannon et al. 2021), and StarDist(Schmidt et al. 2018), Watered(Kornilov and Safonov 2018), utilizing spatial transcriptome staining images as the evaluation medium. Experimental results(Fig. 3 and Fig. 4) indicate that the pre-trained model "cyto2" within the Cellpose framework is the most suitable choice for the whole framework. Cellpose is built upon the architecture of convolutional neural networks and U-Net, it performs high-quality segmentation and classification of different types of cell images. The model can accurately identify the majority of cell nuclei in spatial transcriptome staining images, thereby enhancing the reliability of subsequent analyses in spatial transcriptomics.

Post processing

The selection of positive and negative samples in self-supervised learning is crucial to the model. To mitigate the noise generated by the Cellpose and to ensure the acquisition of more reliable positive samples, we employed a dynamic adaptive morphological erosion algorithm to optimize the cell nucleus boundaries generated by the segmentation model. Through multi-scale iterative erosion operations, we progressively eliminated edge artifacts and discrete noise points, while excluding overly small or irregularly shaped areas resulting from erosion. Ultimately, regions with intact morphology and clear gene expression signals were identified, providing a high-confidence training data foundation for subsequent self-supervised learning. The pseudocode is shown in Algorithm 1.

Algorithm 1: Cell Contour Erosion

Require: $X = n \times matrix$, cell contour maxrix for n cells
M: a rotation matrix of a random angle
$f_{warpaffine}$: affine transformation function
Ensure: E: cell contour matrix after erosion
1: **for** i in iterator$(X_1, X_2, ...X_i)$ **do**
2: $S = F_{warpaffine}(\mathrm{M}, X_i)$
3: $E(X_i, S)(x, y) = \min_{(s_x, s_y) \in S} I\left(x + s_x, y + s_y\right)$
4: **end for**

Cell correction

For each labeled nucleus region we calculate its centroid. The centroid coordinates represent the central point of the nucleus. This can be achieved by computing the average of the x-coordinates and y-coordinates of all the N pixels

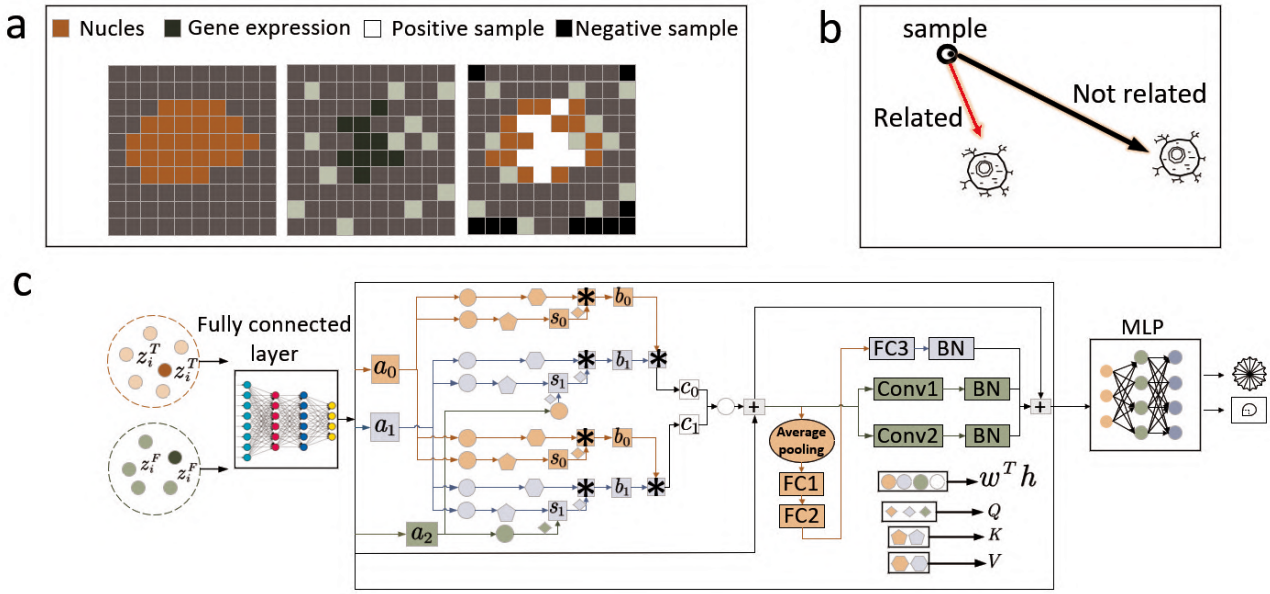

Figure 1: **SSL-CST framework. a**, It indicates that the orange region outlines the nucleus, while the green area represents gene expression information. Positive samples are identified within the nucleus where gene expression is strong, depicted in white, whereas negative samples are those located far from the nucleus with little to no gene expression. **b**, When a sample is positioned in proximity to two adjacent cells, it is more likely to be associated with the closer cell. **c**, Self-supervised learning network architecture. The weight of each cell point expression is adaptively learned through the Transformer-based multi-head self-attention mechanism. See Cell correction for details.

within the nucleus region:

$$C_i(x,y) = \left[\frac{\sum_{j=1}^{N} \text{nucleus\_x}[i][j]}{N}, \frac{\sum_{j=1}^{N} \text{nucleus\_y}[i][j]}{N}\right] \quad (1)$$

The proximity of a point to the center of mass of a nucleus serves as evidence of its association with the cell. The closer a point is to the nucleus's center of mass, the stronger the indication that it belongs to the cell. For each sample point outside the nucleus, calculate the euclidean distance to the centroid of each nucleus and normalize the distance using the exponential function:

$$d_{i,j}(x,y) = \sqrt{(x_i - x_{c_j})^2 + (y_i - y_{c_j})^2} \quad (2)$$

$$d_i = min(d_{i_1}, d_{i_2}, \ldots, d_{i_n}) \quad (3)$$

Upon obtaining the nucleus region, SSL-CST employs a self-supervised learning algorithm to delineate the cytoplasm region(Fig. 1). To ensure the selection of accurate samples for cell identification, a specific criterion is used to define positive and negative samples.

Positive Samples Points that exhibit gene expression within the nucleus are selected after applying image erosion. This selection criterion ensures that only those points indicative of active gene expression are considered as positive samples. Conversely, points located within the nucleus that do not show gene expression are excluded from this category.

Negative Samples The points that are not located within a nucleus, exhibit a distance to the nearest nucleus that exceeds a predefined threshold, and demonstrate no gene expression are selected as negative samples.

The nearest neighbor distance in negative samples is subjected to exponential normalization, while the positive sample is also normalized. The coordinate information of the sample is also subjected to normalization:

$$p_i = \begin{bmatrix} \frac{x_i - \min_x}{\max_x - \min x} \\ \frac{y_i - \min_y}{\max y_y - \min y} \end{bmatrix} \quad (4)$$

To forecast a point's gradient direction into its cell and whether it is within or outside the cell, we employ the following classifier. By taking gene expression information(g), position information(p) and distance information(d) as input, the encoded tensor is obtained through a process involving linear projection and encoding using three fully connected layers:

$$t = gW_1 + pW_2 + dW_3 \quad (5)$$

$W_1 \in R^{N\times D}$,$W_2 \in R^{2\times D}$,$W_3 \in R^D$ represent the weight matrices of the three dense layers.

To facilitate feature extraction, we employ an encoder composed of eight identical modules. Each module comprises two essential components: a self-attention(Vaswani et al. 2017) mechanism and a RMLP. The self-attention mechanism plays a crucial role in capturing dependencies and relationships among different elements within the input data. It makes it possible for the model to provide distinct

segments of the input sequence varying degrees of priority, facilitating efficient feature extraction:

$$[q, k, v] = rU_{qkv} \tag{6}$$

$$A = softmax(\frac{qk^T}{\sqrt{D}}) \tag{7}$$

$$SA(r) = Av \tag{8}$$

In the self-attention architecture, the tensor U_{qkv} is responsible for projecting each point representation into a three-dimensional vector. The dot product between each query and all keys is calculated, and it is then scaled by the square root of D to yield the attention score A. These dot products are passed through the softmax function, resulting in attention scores that represent the weights assigned to neighboring points. These attention scores are then used to weight the corresponding values v, resulting in a D-dimensional weighted representation for each point. Through the use of this process, the model is able to concentrate on pertinent neighboring points and include their data into the representation of each individual point.

In the second module of the architecture, the representation obtained from the attention mechanism undergoes further transformation. This module consists of a stack of multiple RMLP layers. The RMLP layers employ a unique approach where the reparameterization matrix is randomly initialized during each training iteration. This reparameterization matrix is then applied as a linear transformation to the input tensor during forward propagation:

$$RMLP(r) = I \times W \tag{9}$$

This process of random initialization and change can be seen as a noise injection mechanism, which aids in enhancing the model's generalization ability. Within the RMLP layers, the reparameterized matrices are trainable, allowing the model to learn and adjust the transformation based on the given task and data. During each forward propagation, the input tensor is multiplied by the reparameterized matrix through matrix multiplication. The resulting tensor is then processed through the GELU activation function, which introduces nonlinearity and helps capture complex patterns and relationships. By incorporating these RMLP layers, the architecture enables the model to exploit the benefits of noise injection and adaptive linear transformations. This mechanism enhances the model's capacity to generalize well and learn meaningful representations from the input data.

In each encoder layer, the input tensor undergoes layer normalization before being processed by the subsequent block. Layer normalization normalizes the values along each feature dimension of the input tensor, ensuring that the distribution of values remains consistent across different samples. In the architecture, each encoder layer takes the output r_{l-1} of the previous layer as input and generates the output r_l for the subsequent layer. This sequential process continues until we reach the last layer of the encoder:

$$r'_{l-1} = SA(LN(r_{l-1})) + r_{l-1} \tag{10}$$

$$r_l = RMLP(LN(r'_{l-1})) + r'_{l-1} \tag{11}$$

Establish point-to-nucleus direction labels for selected training samples, a whole circle is divided evenly into 16 direction classes:

$$dir = \begin{cases} \text{floor}\left(\frac{\arctan(\frac{y_c - y}{x_c - x})}{2\pi} \times 16\right) & \text{if } y_c - y >= 0 \text{ and } x_c - x >= 0 \\ \text{floor}\left(\frac{\arctan(\frac{y_c - y}{x_c - x}) + \pi}{2\pi} \times 16\right) & \text{if } x_c - x < 0 \\ \text{floor}\left(\frac{\arctan(\frac{y_c - y}{x_c - x}) + 2\pi}{2\pi} \times 16\right) & \text{if } y_c - y < 0 \text{ and } x_c - x >= 0 \end{cases} \tag{12}$$

Points in close proximity to the nucleus tend to have more predictable gradients. We adjust direction prediction by adjusting the prior probability of direction classification based on prior knowledge such as nucleus coordinates and size. Based on the preset cell radius r and the standard deviation per unit cell σ, where the standard deviation is one quarter of the radius. Different expected radii can be set for different parts of the dataset. If the coordinates of the surrounding cell nuclei are not known, we make the assumption that the prior probability of the gradient direction of a point d_g is uniformly distributed:

$$P_{old}(d_g = i) = \frac{1}{16}, for\ i \in [0, 15] \tag{13}$$

Once the surrounding nuclear coordinates of a point are known, the prediction options for the direction are narrowed down to the region where the center of the nucleus can be found. With this additional information, we can restrict the possible directions for the point's gradient to a specific area:

$$P_{new}(d_g = i) = \begin{cases} 0 \text{ No nucleus within } r + \sigma \text{ in the area} \\ \frac{1}{n} \text{ Nucleus within } r + \sigma \text{ in the area} \end{cases} \tag{14}$$

Provided that there are n possible directions containing the nucleus, the predicted probability of the point in the i direction can be calculated as follows:

$$P_{new}(d_g = i|x, s) = \frac{P_{old}(d_g = i|x, s)P_{new}(d_g = i)/P_{old}(d_g = i)}{\sum_{k'=0}^{n} P_{old}(d_g = i|x, s)P_{new}(d_g = i)/P_{old}(d_g = i)} \tag{15}$$

The pseudocode of the algorithm is shown in Algorithm 2. Take the direction class with the highest probability and convert it back to a two-dimensional direction vector. These gradients represent the direction and magnitude of the signal associated with each point. The gradient flow tracking algorithm(Li et al. 2007) is then employed to perform unit segmentation. In the gradient flow tracking algorithm, the direction vectors act as guides for the flow of information. The vectors flow towards a receiver, which corresponds to the center of the nucleus of each cell. Noise present in the orientation vector of the sequenced section or in spots where no RNA is detected can lead to unexpected termination of the flow trace. Therefore, we smooth point-level predictions by averaging the direction vectors and target probabilities of each point and the eight nearest neighbors of the point itself.

Algorithm 2: Main Learning Algorithm

Require: $x \in X$: training samples, each sample x contains three inputs: gene expression (x_g), position (x_p), and distance (x_d).

Ensure: $(y, y_a) \in Y$: y binary classification probability and angle.

for sampled minibatch $\{\boldsymbol{x}_k\}_{k=1}^{N}$ **do**
  $v = W_1 \times x_g + W_2 \times x_p + W_3 \times x_d + b$
  $head_i = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}$
  MultiHead$(Q, K, V) =$
    Concat$(head_1, head_2, \ldots, head_h) \times W$
  for i to n **do**
    $v_1 = \text{MultiHead}(\text{norm}(v))$
    $v_2 = \text{Add}(v_1, v)$
    for units to hidden_units **do**
      $v_3 = v_2 \times W_{\text{units}}$
      $v_4 = c \times v_3 \times \left(1 + \text{erf}\left(\frac{v_4}{\sqrt{2}}\right)\right)$
      $\mathbf{v} = \begin{cases} \frac{x_i}{1-p} & \text{with probability } (1-p) \\ 0 & \text{with probability } p \end{cases}$
    end for
    return v
  end for
end for
Define $\mathcal{L}_y(y, \hat{y}) = -y \log(\hat{y}) - (1-y)\log(1-\hat{y})$
Update networks W and b to minimize $\mathcal{L}$

Experiments

Experimental Setup All experiments conducted in the study were performed on a server equipped with an eight-core Intel(R) Xeon(R) Gold 6240 CPU and an NVIDIA A100-SXM4-80GB GPU. All experiments use the Adam optimizer unless otherwise stated. The initial learning rate is set to $1e^{-4}$. The batchsize is set to 256. To prevent overfitting and enhance training efficiency, we established a maximum of 200 epochs and implemented an early stopping mechanism: training was automatically terminated if the validation loss did not decrease for 10 consecutive epochs.

Datasets We conducted the evaluation of the proposed cell segmentation method on three spatial transcriptomics datasets obtained by different techniques: Stereo-seq, SeqFISH+ and Merscope-Ovarian.

SeqFISH+ (Eng et al. 2019): The dataset comprises single decoded mRNA point positions obtained from experiments conducted on NIH/3T3 cells, encompassing a total of 10,000 genes. Notably, the database includes the true expression values for each individual cell.

Stereo-seq (Chen et al. 2022): The dataset is a three-dimensional sequencing data of mouse embryos. encompassing a significantly greater number of genes compared to other datasets. The specific specimen utilized in this study is E14.5E1S3, which contains 27,413 genes and 4,873 cells.

Merscope-Ovarian (Huang et al. 2024): The Merscope-Ovarian dataset comprises spatial transcriptomics data derived from pathological sections of human ovarian cancer.

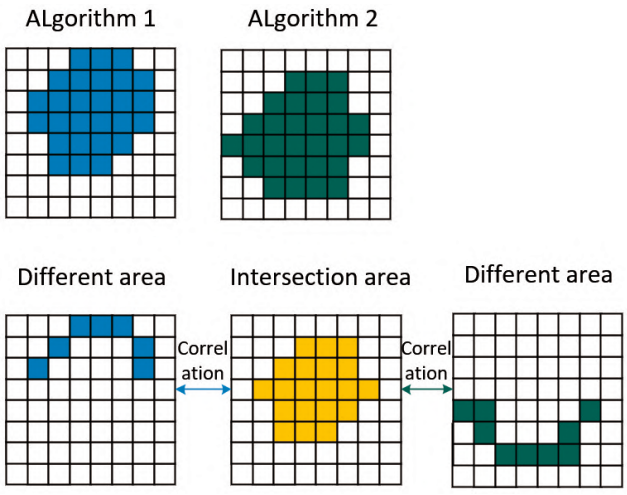

Figure 2: The intersecting area, depicted in yellow, represents the region where the cells identified by both methods overlap. The green and blue areas correspond to the disjoint regions identified by each method individually. Calculate the correlation between gene expression levels in the disjoint area and the intersecting area.

For our analysis, we specifically utilized data from the third section, which includes 500 sequenced genes and 71,381 cells.

Evaluation indicators In the datasets obtained from the Stereo-seq and Merscope-Ovarian methods, no ground truth of cell segmentation was provided. To evaluate these datasets, Viktor Petukhov(Petukhov et al. 2022) proposed a common method for evaluating cells generated by different segmentation methods. In this evaluation method, the segmentation results from various algorithms are compared to derive their intersection and difference areas. The intersection area typically corresponds to the central region of the cell, which serves as a critical reference point for assessing segmentation accuracy(Fig. 2). The correlation between gene expression levels in each difference area and those in the intersection area is calculated using the Pearson correlation coefficient. A stronger correlation indicates a greater similarity between the gene expression profiles in the difference areas and those in the intersection area, the more precise the cell segmentation achieved by the respective method.

Experimental Results The number of nucleus segmented by four distinct segmentation methods(C-cellpose, W-water, D-deepcell, S-stardist) was assessed in both the Stereo-seq and SeqFISH+ datasets(Fig. 3). Cellpose achieved the highest segmentation count of nuclei in both datasets(The number of cell nuclei in the Stereo-seq dataset was 4126 and in the SeqFISH+ dataset this number was 857). Further analysis revealed the watershed algorithm shows a large area of nucleus adhesion, while the startdist algorithm shows varying degrees of nucleus fragmentation(Fig. 4). This observation further substantiates the conclusion that the watershed

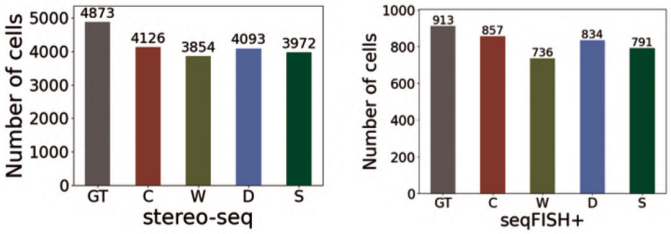

Figure 3: Comparison of the number of nuclei segmented by different algorithms in Stereo-seq and Seqfish+ with ground truth. The x-axis represents the different segmentation methods, while the y-axis indicates the number of cells.

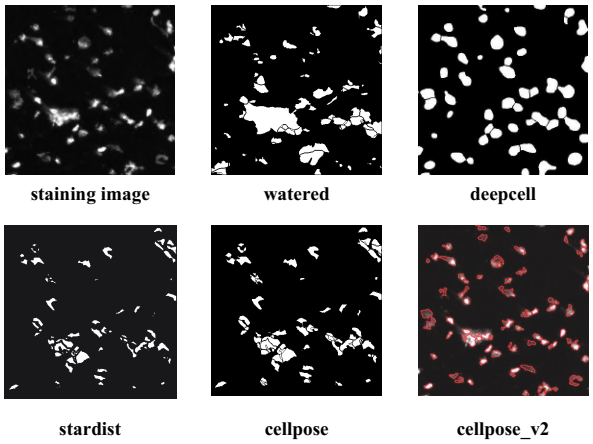

Figure 4: The segmentation results of various algorithms applied to nucleus staining images.

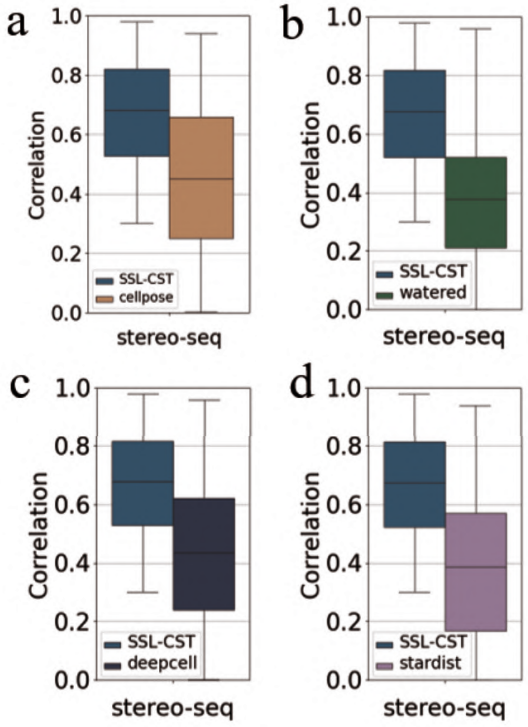

Figure 5: Correlation of SSL-CST with other methods on Stereo-seq datasets. The correlation between SSL-CST and Cellpose is observed to be the highest, indicating a strong alignment between the segmentation results produced by these two methods. This can be attributed to the fact that SSL-CST leverages the foundational segmentation results of cell nuclei generated by Cellpose as its starting point.

exhibits a low correlation with other segmentation methods. The transcriptome data obtained through the watershed algorithm often do not correspond to the same cell. Although the DeepCell algorithm can more accurately identify the outline of the nucleus, some nucleus are still not identified.

In the analysis of the Stereo-seq dataset, the average correlation coefficient of the SSL-CST segmentation results is found to be 23% higher than that of Cellpose, demonstrating a marked improvement over other segmentation methods(Fig. 5). The increased correlation underscores the capability of SSL-CST to accurately delineate the cytoplasm region.

We measured the cell diameters obtained through SSL-CST segmentation and compared them with the diameters of cell nuclei segmented by other methods(Fig. 6). The results indicated that SSL-CST provided a more realistic estimation of cell diameter, consistently yielding values that exceeded those of any cell nucleus diameter measured by the alternative segmentation techniques. This finding indicates that SSL-CST effectively integrates the cytoplasm into cell segmentation results, enhancing the overall accuracy of the segmentation process.

In the SeqFISH+ dataset, the manually annotated cell segmentation ground truth can be obtained from the original dataset. We used the Intersection over Union (IoU)(Rezatofighi et al. 2019) metric to evaluate the consistency of cell segmentation with the ground truth when assessing the performance of several algorithms on the seqFISH dataset. The results (Fig. 7) of the study indicate that the cell segmentation method performs exceptionally well in the SeqFISH+ dataset, surpassing other segmentation methods in terms of accuracy.

Furthermore, we conducted a comparative analysis of the correlation between the SSL-CST method and the other three segmentation methods using the Merscope-Ovarian dataset(Fig. 8). The results indicated that SSL-CST achieved an average correlation that was over 4% higher than that of the SCS method and exhibited an impressive increase in correlation compared to the other two methods. One noteworthy phenomenon observed is that when transcriptome throughput is low, the segmentation results obtained from various methods do not perform as well as those derived from high-throughput stereo datasets. This discrepancy suggests that lower throughput may limit the amount of available transcriptomic information, potentially affecting the accuracy and reliability of cell segmentation.

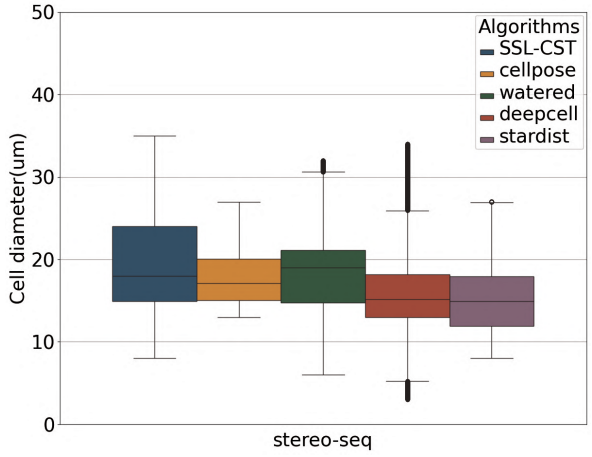

Figure 6: Comparison of nuclei diameters by different methods(Cellpose, Watered, DeepCell, StarDist) and cell diameters segmented by SSL-CST on the Stereo-seq dataset.

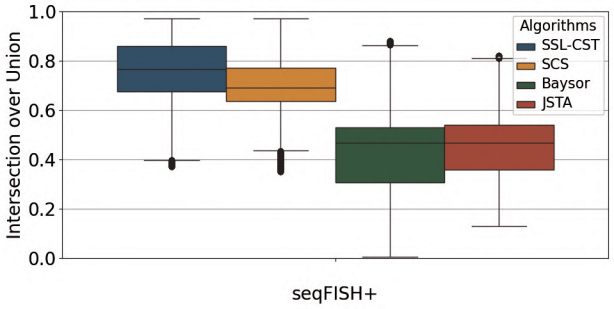

Figure 7: In the case of the cell segmentation method applied to the seqFISH dataset, the average IoU score of SSL-CST across all cells was found to be 0.772. SCS achieves an average of 0.687, Baysor scores 0.495 and JSTA scores 0.496. This enhanced performance underscores the potential of SSL-CST for reliable cell segmentation.

Ablation study We conducted ablation experiments utilizing the SeqFISH+ dataset, which includes true values, to assess the impact of specific features on model performance(Fig. 9). In these experiments, we systematically removed the distance feature and the position feature from the model during training. Upon training the model for 200 epochs, the removal of the distance feature resulted in a decrease in accuracy of nearly 20% and the removal of the position feature led to a reduction in accuracy of approximately 10%. These results highlight the significance of both the distance and position features in enhancing the model's performance, underscoring their contributions to the overall accuracy of the segmentation process.

Conclusion

In this study, we proposed a novel cell segmentation method based on the principles of self-supervised learning. A unique property is that the method fully integrates the information from staining images and gene expression data, leveraging both the positional characteristics and gene expression profiles of each point within the tissue sample. A self-supervised network is employed to classify all points in the stained image to obtain high-precision cells. Through extensive testing on multiple datasets, the method outperforms other state-of-the-art methods, demonstrating superior segmentation performance. Given the substantial computational demands associated with integrating high-dimensional transcriptomic and imaging data, we plan to further optimize the algorithm's runtime and memory usage in future work. This will enable the development of practical solutions for large-scale multimodal biomedical data analysis. Furthermore, incorporating domain-specific knowledge and prior information into the segmentation framework is worth exploring. Leveraging prior knowledge about cell morphology, tissue structure, or specific staining patterns can guide the segmentation process and improve its accuracy and efficiency.

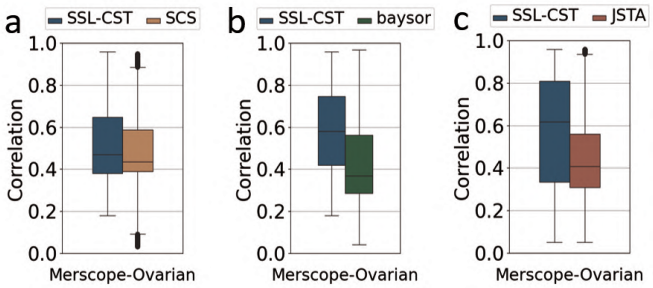

Figure 8: A comparative analysis of the correlation between SSL-CST and the segmentation methods SCS, Baysor, and JSTA was conducted using the Merscope-Ovarian dataset. The results reveal that the correlation coefficient for SSL-CST is slightly higher than that of SCS, indicating a marginal improvement in segmentation accuracy. In contrast, SSL-CST exhibits a significantly higher correlation when compared to both Baysor and JSTA.

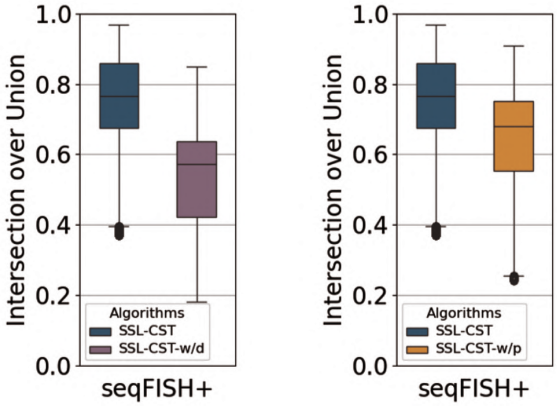

Figure 9: Testing of different options of model parts. Figure a showcases the result of utilizing the watershed method as a replacement for the Cellpose model in the segmentation process. Figure b demonstrates the outcome of removing the distance feature engineering from the analysis.

Acknowledgments

This study is supported by National Natural Science Foundation of China (Grants No. 62572157 and 62462022) and Hainan University Research Fund (Grant KYQD(ZR)-21079).

References

Asp, M.; Bergenstråhle, J.; and Lundeberg, J. 2020. Spatially resolved transcriptomes—next generation tools for tissue exploration. *BioEssays*, 42(10): 1900221.

Bannon, D.; Moen, E.; Schwartz, M.; Borba, E.; Kudo, T.; Greenwald, N.; Vijayakumar, V.; Chang, B.; Pao, E.; Osterman, E.; et al. 2021. DeepCell Kiosk: scaling deep learning–enabled cellular image analysis with Kubernetes. *Nature methods*, 18(1): 43–45.

Bhang, H.-e. C.; Ruddy, D. A.; Krishnamurthy Radhakrishna, V.; Caushi, J. X.; Zhao, R.; Hims, M. M.; Singh, A. P.; Kao, I.; Rakiec, D.; Shaw, P.; et al. 2015. Studying clonal dynamics in response to cancer therapy using high-complexity barcoding. *Nature medicine*, 21(5): 440–448.

Bray, F.; Laversanne, M.; Sung, H.; Ferlay, J.; Siegel, R. L.; Soerjomataram, I.; and Jemal, A. 2024. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 74(3): 229–263.

Chen, A.; Liao, S.; Cheng, M.; Ma, K.; Wu, L.; Lai, Y.; Qiu, X.; Yang, J.; Xu, J.; Hao, S.; et al. 2022. Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. *Cell*, 185(10): 1777–1792.

Chen, H.; Li, D.; and Bar-Joseph, Z. 2023. SCS: cell segmentation for high-resolution spatial transcriptomics. *Nature methods*, 20(8): 1237–1243.

Chhikara, B. S.; Parang, K.; et al. 2023. Global Cancer Statistics 2022: the trends projection analysis. *Chemical Biology Letters*, 10(1): 451–451.

Dagogo-Jack, I.; and Shaw, A. T. 2018. Tumour heterogeneity and resistance to cancer therapies. *Nature reviews Clinical oncology*, 15(2): 81–94.

Eng, C.-H. L.; Lawson, M.; Zhu, Q.; Dries, R.; Koulena, N.; Takei, Y.; Yun, J.; Cronin, C.; Karp, C.; Yuan, G.-C.; et al. 2019. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature*, 568(7751): 235–239.

Gohil, S. H.; Iorgulescu, J. B.; Braun, D. A.; Keskin, D. B.; and Livak, K. J. 2021. Applying high-dimensional single-cell technologies to the analysis of cancer immunotherapy. *Nature Reviews Clinical Oncology*, 18(4): 244–256.

He, Y.; Tang, X.; Huang, J.; Ren, J.; Zhou, H.; Chen, K.; Liu, A.; Shi, H.; Lin, Z.; Li, Q.; et al. 2021. ClusterMap for multi-scale clustering analysis of spatial gene expression. *Nature communications*, 12(1): 5909.

Huang, R.; Kratka, C. E.; Pea, J.; McCann, C.; Nelson, J.; Bryan, J. P.; Zhou, L. T.; Russo, D. D.; Zaniker, E. J.; Gandhi, A. H.; et al. 2024. Single-cell and spatiotemporal profile of ovulation in the mouse ovary. *bioRxiv*, 2024–05.

Kornilov, A. S.; and Safonov, I. V. 2018. An overview of watershed algorithm implementations in open source libraries. *Journal of Imaging*, 4(10): 123.

Li, G.; Liu, T.; Tarokh, A.; Nie, J.; Guo, L.; Mara, A.; Holley, S.; and Wong, S. T. 2007. 3D cell nuclei segmentation based on gradient flow tracking. *BMC cell biology*, 8: 1–10.

Littman, R.; Hemminger, Z.; Foreman, R.; Arneson, D.; Zhang, G.; Gómez-Pinilla, F.; Yang, X.; and Wollman, R. 2021. Joint cell segmentation and cell type annotation for spatial transcriptomics. *Molecular systems biology*, 17(6): e10108.

Liu, Y.; Yang, M.; Deng, Y.; Su, G.; Enninful, A.; Guo, C. C.; Tebaldi, T.; Zhang, D.; Kim, D.; Bai, Z.; et al. 2020. High-spatial-resolution multi-omics sequencing via deterministic barcoding in tissue. *Cell*, 183(6): 1665–1681.

Pachitariu, M.; and Stringer, C. 2022. Cellpose 2.0: how to train your own model. *Nature methods*, 19(12): 1634–1641.

Petukhov, V.; Xu, R. J.; Soldatov, R. A.; Cadinu, P.; Khodosevich, K.; Moffitt, J. R.; and Kharchenko, P. V. 2022. Cell segmentation in imaging-based spatial transcriptomics. *Nature biotechnology*, 40(3): 345–354.

Prabhakaran, S. 2022. Sparcle: assigning transcripts to cells in multiplexed images. *Bioinformatics advances*, 2(1): vbac048.

Racle, J.; de Jonge, K.; Baumgaertner, P.; Speiser, D. E.; and Gfeller, D. 2017. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *elife*, 6: e26476.

Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; and Savarese, S. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 658–666.

Schmidt, U.; Weigert, M.; Broaddus, C.; and Myers, G. 2018. Cell detection with star-convex polygons. In *Medical image computing and computer assisted intervention–MICCAI 2018: 21st international conference, Granada, Spain, September 16-20, 2018, proceedings, part II 11*, 265–273. Springer.

Sung, H.; Ferlay, J.; Siegel, R. L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; and Bray, F. 2021. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 71(3): 209–249.

Tellez-Gabriel, M.; Ory, B.; Lamoureux, F.; Heymann, M.-F.; and Heymann, D. 2016. Tumour heterogeneity: the key advantages of single-cell analysis. *International journal of molecular sciences*, 17(12): 2142.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, X.; Park, J.; Susztak, K.; Zhang, N. R.; and Li, M. 2019. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nature communications*, 10(1): 380.

Zheng, B.; and Fang, L. 2022. Spatially resolved transcriptomics provide a new method for cancer research. *Journal of Experimental & Clinical Cancer Research*, 41(1): 179.

Zhuang, X. 2021. Spatially resolved single-cell genomics and transcriptomics by imaging. *Nature methods*, 18(1): 18–22.