Vispro improves imaging analysis for Visium spatial transcriptomics

Huifang Ma[1], Yilong Qu[1], Anru Zhang[1], and Zhicheng Ji[1,†]

[1]Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC, USA.
[†]Corresponding author. E-mail: zhicheng.ji@duke.edu

ABSTRACT

Spatial transcriptomics (ST) is a cutting-edge technology that enables the comprehensive analysis of gene expression while preserving the spatial context of tissues. Within this framework, histological images serve a crucial role, providing spatially cohesive information that is often challenging to capture through gene expression alone. However, the imaging process in ST data presents inherent challenges, such as variability in image quality, artifacts from fiducial markers, and difficulty in distinguishing tissue regions. These challenges can significantly impair the accuracy and effectiveness of downstream analyses. To address these limitations, we developed Vispro, an end-to-end, fully automated image processing tool tailored for ST data. Vispro integrates four key modules: fiducial marker detection, marker removal and image restoration, tissue region detection, and segmentation of disconnected tissue areas. These modules systematically enhance the quality of ST images, reducing artifacts and improving tissue segmentation accuracy. Furthermore, by improving the integrity of image data, we demonstrate that Vispro facilitates downstream analyses such as cell segmentation and image registration, empowering researchers to extract more detailed and accurate biological insights from complex tissue architectures.

Introductions

Spatial transcriptomics (ST) is a transformative technology that simultaneously measures the gene expression profiles, spatial locations, and imaging information of the same set of cells. Compared to single-cell sequencing that cannot capture the spatial and imaging information, ST provides deeper understanding of how different cell types are spatially organized within a tissue and allows one to study the morphology of cells. The imaging information accompanying ST has become indispensable in many analytical methods[1–11] designed for ST data, with two primary approaches currently being explored. One approach[1–9] leverages image data to predict gene expression or facilitate gene analysis, aiming to extract molecular insights from histological images. A notable example is TESLA[9], which estimates tissue contours and generates superpixel regions based on histological similarity, enabling high-resolution gene imputation and improving outcomes in cancer ecosystem analysis. Others seek to establish a direct correlation between histological images and gene expression, using this relationship to predict gene expression on widely accessible and cost-effective H&E images. The other approach[10,11] integrates both gene expression data and imaging information to develop more comprehensive models, using the spatial correlation between the two modalities to improve the accuracy and depth of biological interpretation. For instance, SpaGCN[10] integrates histology with gene expression and the spatial locations of cells to infer spatial domains and identify spatially variable genes. Similarly, stLearn[11] combines imaging, gene expression, and spatial distances between cells to map spatiotemporal trajectories and cell–cell interactions. Additionally, images have shown great potential in improving the accuracy of aligning multiple tissue sections, as highlighted by several studies[12]. This alignment facilitates integrative analysis of ST data across different samples, providing deeper insights into spatial gene expression patterns and tissue architecture.

Despite the promise of image-based approaches in ST, multiple challenges arise when analyzing images from ST datasets. First, images from 10x Visium, arguably the most popular ST platform, contain artificial reference points, or fiducial markers (Figure 1a,b), intended to detect tissue regions. Although designed to outline tissue areas, fiducial markers often present challenges during image processing. They frequently overlap with tissue regions due to operational errors, obscuring critical structures and textures. Additionally, their blob-like shape often leads to misclassification as tissue structures by many current methods. Furthermore, their consistent global square arrangement can mislead models attempting to learn the spatial patterns of tissue, resulting in inaccurate interpretation of global tissue architecture. Second, most ST images consist of both tissue regions, where cellular structures are visible, and blank regions, referred to as background (Figure 1a,c). The background often contains residual stains and random noise, which blur tissue boundaries and severely interfere when complete images are used for predicting gene expression. Third, a single ST image may capture multiple tissue samples, especially when using tissue microarrays (TMAs), collected from different spatial locations or individuals (Figure 1a,d). Without proper segmentation of these distinct regions, downstream analysis becomes significantly complicated. The failure to separate these regions impedes

accurate spatial correlation modeling, as tissue morphology from different locations can confound one another. This challenge obscures the heterogeneity between samples, making it difficult to interpret biological differences and compromising the accuracy of spatial analyses.

These challenges have been largely overlooked and underexplored in the existing literature, especially for the handling of fiducial markers. The 10x Genomics Space Ranger[13] and Loupe Browser[14] can automatically or manually detect fiducial markers. However, they rely on a standardized fiducial marker template for alignment, assuming that all markers are perfectly circular and uniformly arranged. This assumption fails to account for the deformations or misprints commonly present in real images (Figure 1a,e). More critically, these tools are unable to automatically remove fiducial markers from images or recover distorted tissue regions where the markers overlap with tissue, a step essential for accurately analyzing tissue similarity. Besides, existing methods for detecting and segmenting tissue regions are ineffective when applied directly to ST images without first removing fiducial markers, leading to erroneous results, as we will demonstrate. Although manual manipulation of images can address these issues, it is time-consuming and impractical for large-scale datasets. In summary, there lacks a method that can automatically and efficiently process ST images to fully remove the technical artefacts and produce images that can be readily used in downstream analysis.

To address this issue, we developed Vispro, a fully automated image processing tool for ST images (Figure 1f). Vispro includes four sequential modules: detecting fiducial markers, removing these markers and restoring the image, detecting tissue regions, and segregating disconnected tissue regions. The tool outputs cleaned images with accurately defined tissue regions, ready for integration into other image analysis software. We demonstrate that Vispro provides higher-quality image information in ST, including more accurate fiducial marker localization and tissue boundary detection. Furthermore, we show that images processed by Vispro significantly improve the performance of downstream tasks, such as cell segmentation and image registration. Thus, Vispro serves as a valuable tool for optimizing the use of image data in ST.

Results

Method overview

We developed a customized deep-learning framework with built-in modules for fiducial marker detection, fiducial marker recovery, tissue segmentation, and fragment segregation in ST images. The fiducial marker detection module uses a deep neural network based on the traditional U-Net architecture, with tailored layer parameters and loss functions designed specifically for this task. To train the model, we manually annotated approximately 100 thousand training images on 167 ST images. The network outputs a binary segmentation mask of fiducial markers. Once the marker is detected as a binary segmentation mask, both the mask and the original image are fed into an image inpainting module to restore the content in the marker area while preserving the image outside of it. Following restoration, a U^2Net neural network is applied to detect the salient object and simultaneously separate the tissue from the background. Both models are pre-trained on large natural image datasets to ensure effectiveness. Finally, disjoint tissue regions are identified using a brute-force search of pixel connectivity with the OpenCV library.

Vispro accurately identifies and removes fiducial markers

We first compared the performance of Vispro with the standard 10x Genomics pipeline for detecting and removing fiducial markers in a cross-validation study. Figure 2 a shows fiducial markers identified by both methods in a typical example. In this image, the fiducial markers exhibit severe deformation, leading to a disordered arrangement that significantly deviates from their original design and placement. As a result, the 10x template struggles to align with the markers, leaving substantial portions of the marker areas uncovered. In contrast, our method, which is learning-based, infers texture information and accurately locates the marker areas, even in the presence of such distortions. Figure 2b shows a quantitative measure of intersect over union (IoU) comparing the results of the two methods with the gold standard. Vispro substantially improves the IoU of 10x Genomics pipeline by 11.09%, showing that it can more accurately identify the fiducial markers than the existing method.

Figure 2c further shows the results of inpainting the identified fiducial markers. The binary masks of fiducial markers obtained by the two methods were used as inputs to the same inpainting process. When using the fiducial markers identified by Vispro, the inpainting process successfully removes all fiducial markers from the images. In contrast, noticeable fiducial markers remain on the images when using the 10x pipeline's results for inpainting. This demonstrates the superior accuracy of fiducial markers identified by Vispro and highlights how the precision of fiducial marker identification significantly impacts downstream image processing.

Vispro accurately identifies tissue regions

After inpainting fiducial markers identified by Vispro, we next evaluated the performance of Vispro in differentiating tissue regions from background regions (Figure 3) and a subsequent module of segregating disconnected tissue regions (Figure 4). The

computationally identified tissue regions and segregated tissue subregions were quantitatively compared to the gold standard of manually annotated tissue regions, using IoU as the metric. As a comparison, we also included results using the original images.

Vispro is able to accurately identify the tissue regions from the background regions (Figure 3). The IoU metrics are close to 1 in most cases, suggesting that the tissue regions identified by Vispro match with the gold standard almost perfectly. In comparison, directly identifying tissue regions without inpainting fiducial markers results in substantially worse results, with Iou as low as 0.55. The presence of fiducial markers leads to inaccurately enlarged or truncated tissue regions, especially when fiducial markers overlap with actual tissue regions.

Vispro effectively identifies disconnected tissue subregions (Figure 4). In most cases, it accurately detects the correct number of tissue subregions, with the exception of more complex cases involving over 10 fragmented tissue sections, as seen in the third row of Figure 4. However, the Intersection over Union (IoU) metric remains high, indicating that Vispro consistently predicts the tissue area with precision. In contrast, segmentation based solely on the original images shows a significant decline in performance, particularly by missing key tissue regions or incorrectly merging thin structures with the background, resulting in substantially lower IoU values. Additionally, background noise, along with the presence of fiducial markers, sometimes causes errors such as merging distinct tissue regions or cutting off tissue areas.

Vispro improves cell segmentation

Cell segmentation aims to identify and separate individual cells in an image[15]. Accurately identifying the locations of cells is crucial for understanding the spatial organization of different cell types within a tissue. Since most cell segmentation methods purely rely on imaging information, the quality of images will largely affect the performance of cell segmentation.

We applied StarDist[16], a commonly used cell segmentation method, to images processed by the whole Vispro pipeline, images that only went through the fiducial marker removal step of Vispro, or the original images (Figure 5). Since the goal of Vispro is to remove the unwanted areas outside tissue regions, we only evaluated if the cell segmentation method detects any cells in regions outside of the tissue, where no cell is expected to exist. These cells are treated as false positives. The accuracy of cell segmentation within tissue regions is largely affected by the performance of the segmentation algorithm itself and is out of scope of this study.

StarDist identifies thousands of false positive cells when using the original images or images that only went through fiducial marker removal. A main reason is that the segmentation method falsely treats fiducial markers and background noise as images of cells, leading to many false positives. In comparison, the numbers of false positive cells reduce substantially when images processed with Vispro were used for segmentation, demonstrating the necessity of removing unwanted technical artifacts from the images. These results suggest that Vispro improves the performance of cell segmentation and leads to more accurate characterization of cells' spatial locations.

Vispro improves image registration

Image registration is the operation of transforming a source image into the same coordinate system of a target image so that the two images are directly comparable. Many studies generate multiple ST samples from similar tissue types[17]. Registration these images to a common coordinate system allows researchers to compare spatial and imaging features across individuals. Similar to cell segmentation, image registration relies solely on imaging information and its performance will be affected by the quality of the images.

We tested the performance of bUnwarpJ[18], a state-of-the-art image registration method, using images processed by the whole Vispro pipeline or the original images (Figure 6). We considered two settings of image registration. In the first setting (Figure 6a), both the source image and the target image were generated by the Visium workflow and include fiducial markers. In the second setting (Figure 6b), the source image was generated by the Visium workflow and include fiducial markers, while the target image was generated by the standard H&E histology workflow and does not include fiducial markers. The second setting is useful for Visium CytAssist, an instrument facilitating the transfer of transcriptomic probes from standard glass slides to Visium slides. The performance of image registration was quantitatively evaluated by two numerical measures (Methods) of structural similarity index (SSIM) and mutual information (MI).

Figure 6 demonstrates the performance of image registration in the two settings. For each case, a small part of the registered image was zoomed in to highlight the difference in registration results. When fiducial markers are present, bUnwarpJ tries to align the fiducial markers in addition to the tissue regions. Since the position and layout of fiducial markers are affected by technical factors (e.g., placement of the tissue on the slide), this leads to inferior performance in registering some tissue regions, such as the misaligned tissue structures in the zoomed in regions. In comparison, the performance of image registration improves substantially using images processed by Vispro. With the removal of the fiducial markers, the registration process focuses solely on the tissue regions and the highlighted tissue structures are better aligned. Along with the improved SSIM and MI, these results suggest that Vispro can improve the performance of image registration.

Conclusion

We present Vispro, an automated image processing tool for ST images generated by 10x Visium platform. Vispro provides four consecutive modules to clean the ST images and extract images of tissue regions. Vispro is able to accurately identify the fiducial markers and tissue regions in real examples, outperforming the default processing pipeline provided by 10x. We have shown that processing images with Vispro can substantially improve the performance of downstream imaging analysis tasks, including cell segmentation and image registration.

Methods

Vispro module 1: detecting fiducial markers

Training data collection

We collected 167 datasets generated by 10x Visium platform from 48 studies. The images include cases with severe overlap between tissue and markers as well as properly localized markers. To generate training data, we applied the Hough Transformation[19] for initial circle detection, though many of these were noisy detections. We then manually classified the circles into markers within tissue, markers in the background, and false positives. Additionally, we manually annotated markers missed by the Hough Transformation, creating a comprehensively annotated dataset, yielding 92,685 marker annotations.

Neural Network Architecture

The U-Net is a fully convolutional network with a symmetric architecture consisting of a downsampling path (encoder) and an upsampling path (decoder). Below is the detailed mathematical description of a U-Net with 4 downsampling layers and 3 upsampling layers.

Network input The input to the U-Net is an image represented by a matrix, $X \in \mathbb{R}^{H \times W \times C}$, where H is the height, W is the width, and C is the number of channels (e.g., RGB).

Detailed architecture Downsampling Path (Encoder) The downsampling path consists of repeated applications of convolutional layers followed by max-pooling operations. At each downsampling layer, the spatial resolution is reduced, while the number of feature maps increases. For each downsampling layer, the convolution operation is represented as:

$$X_{\text{conv}}^{(l)} = W^{(l)} * X^{(l-1)} + b^{(l)} \tag{1}$$

where $W^{(l)}$ is the weight matrix for layer l, $*$ denotes the convolution, and $b^{(l)}$ is the bias term. The output of the convolution is passed through a ReLU activation function: $X_{\text{relu}}^{(l)} = \max(0, X_{\text{conv}}^{(l)})$. After each convolution, a max-pooling operation reduces the spatial dimensions by half:$X_{\text{pool}}^{(l)} = \max_{\text{pool}}(X_{\text{relu}}^{(l)})$. This process is repeated for 4 layers, progressively downsampling the image and increasing the number of feature maps.

Bottleneck Layer At the bottom of the U-Net, the bottleneck layer connects the downsampling and upsampling paths. This layer applies a convolution without further downsampling:

$$X_{\text{bottleneck}} = W_{\text{bottleneck}} * X_{\text{pool}}^{(4)} + b_{\text{bottleneck}} \tag{2}$$

Upsampling Path (Decoder) The upsampling path consists of up-convolution (transpose convolution) layers that gradually restore the original image size while combining information from corresponding layers in the downsampling path. The upsampling process is represented by the transpose convolution:

$$X_{\text{upconv}}^{(l)} = W_{\text{upconv}}^{(l)} * X^{(l+1)} + b_{\text{upconv}}^{(l)} \tag{3}$$

where $W_{\text{upconv}}^{(l)}$ is the weight matrix for the transpose convolution, and $X^{(l+1)}$ is the output from the layer above. After each upsampling, the feature map from the corresponding layer in the downsampling path is concatenated:

$$X_{\text{concat}}^{(l)} = \left[X_{\text{conv}}^{(l)}, X_{\text{upconv}}^{(l)}\right] \tag{4}$$

The concatenated feature map is then passed through another convolution to further refine the output: $X_{\text{conv-up}}^{(l)} = W_{\text{conv-up}}^{(l)} * X_{\text{concat}}^{(l)} + b_{\text{conv-up}}^{(l)}$. This process is repeated for 3 layers, progressively increasing the spatial resolution and decreasing the number of feature maps.

Output Layer The final layer of the U-Net is a 1×1 convolution that reduces the number of channels to the desired output,

typically one channel for binary segmentation:

$$X_{\text{output}} = W_{\text{final}} * X^{(1)}_{\text{conv-up}} + b_{\text{final}} \tag{5}$$

The output is passed through a sigmoid activation function for binary segmentation: $X_{\text{final}} = \sigma(X_{\text{output}})$ where $\sigma(X)$ is the sigmoid function:

$$\sigma(X) = \frac{1}{1+e^{-X}} \tag{6}$$

ensuring that the output is in the range $[0,1]$, representing the probability of each pixel belonging to the foreground or background class.

Loss function To address the imbalanced ratio of marker-area with non-marker area, we employed the focal loss[20] as the loss function of the neural network. Focal loss builds upon the conventional cross-entropy loss. It downweighs the loss contributed by easily classifiable classes and upweighs the loss contributed by more challenging classes, leading to improved results compared to commonly used mean square error loss on imbalanced datasets.

The focal loss is defined as:

$$\mathrm{L}(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t)$$

where p_t is the predicted probability for the true class. For positive class (label = 1), $p_t = p$, and for the negative class (label = 0), $p_t = 1-p$, α_t is a balancing factor to address class imbalance, where $\alpha_t = \alpha$ for the positive class and $\alpha_t = 1-\alpha$ for the negative class, γ is the focusing parameter, which adjusts the rate at which easy examples are down-weighted. Higher γ values focus more on hard examples.

Expanding for both positive and negative classes in a binary classification setting, the focal loss can be written as:

$$\text{Focal Loss} = -\alpha(1-p)^{\gamma}y\log(p) - (1-\alpha)p^{\gamma}(1-y)\log(1-p)$$

Where, $y \in \{0,1\}$ is the ground truth label, $p \in [0,1]$ is the predicted probability for the positive class, $\alpha \in [0,1]$ is the balancing factor for the positive class, $\gamma \geq 0$ is the focusing parameter. Given that the fiducial marker area takes a ratio of approximately 0.05 in the collected dataset, we set the alpha parameter of the marker class to 0.95 and the background class to 0.05 to force the model focusing more on the marker area. The γ is empirically set to 2, which works well in real cases.

Network output The output to the U-Net is a binary gray image represented by a matrix, $X \in \mathbb{R}^{H\times W}$, where H is the height, W is the width. In this matrix, the value 1 indicates pixels that belong to the marker area, while 0 indicates pixels that do not belong to the marker area.

Training Procedure

We framed fiducial marker detection as an object segmentation task, generating a binary mask the same size as the input image. One of the key challenges in this task is the severe data imbalance, as fiducial markers occupy only a small fraction of the spatial transcriptomics (ST) image. This imbalance, between positive (marker) and negative (non-marker) areas, required specific strategies to improve learning. To mitigate this, we implemented a sampling layer to adjust the frequency of sample occurrence in the network. We also annotated fiducial markers within and outside tissue regions, allowing us to calculate a slice-wide marker overlap factor, defined as the ratio of in-tissue markers to the total number of markers. Using this factor, we classified images into three categories: hard examples (marker factor > 0.8), moderate examples (0.3 < marker factor $\leq$ 0.8), and easy examples (marker factor < 0.3). To further address the data imbalance, we increased the occurrence of hard samples threefold and moderate samples twofold in the training dataset, ensuring the model is effectively trained on the most challenging examples.

During training, all images underwent random augmentations, including flipping, scaling, rotation, brightness adjustment, and Gaussian blurring, to enhance the robustness of the model. The network was trained extensively over 800 epochs, ensuring robust learning and high performance in fiducial marker segmentation. For model evaluation, we performed leave-one-out cross-validation, splitting our limited dataset into 10 groups. In each iteration, 9 groups were used for training and 1 group for testing. This approach allowed us to perform self-evaluation despite the limited data availability.

Vispro module 2: Fiducial marker inpainting

Image inpainting generally involves filling in missing or corrupted regions of an image by generating visually plausible content that seamlessly blends with the surrounding area. However, most traditional inpainting methods rely on local convolutions, which work well for small missing areas but struggle with larger regions due to their limited receptive field. Although individual

fiducial markers are small, they often form long-range patterns across the image. To address this, we turned to research on Fourier Convolutions[21], which specifically tackles this limitation by operating in the frequency domain. This approach allows the model to efficiently capture long-range dependencies, crucial for filling large missing areas, as it can propagate information over much greater distances compared to spatial convolutions. Additionally, while conventional methods often fail to generalize well across varying resolutions, Fourier Convolutions are designed to be resolution-robust, maintaining consistent performance in both low- and high-resolution images. In this case, the model can be pre-trained on large volumes of natural images, making it particularly effective for the high-resolution images found in spatial transcriptomics (ST). We incorporated the implementation and pre-trained model[22] into our built-in restoration module, seamlessly integrating it to enhance the performance of our system. Specifically, the model input consists of a combined image, represented by a matrix $X \in \mathbb{R}^{H \times W \times C}$, and a corresponding mask, represented by a matrix $M \in \mathbb{R}^{H \times W}$, where H is the height, W is the width, and C is the number of channels (e.g., RGB). In the mask matrix, a value of 1 indicates pixels that need to be filled in, while 0 indicates pixels that can remain unchanged.

Vispro module 3: detecting tissue regions and removing background regions

For the task of tissue segmentation, we adopt the U²-Net model[23] due to its proven effectiveness in handling complex segmentation tasks, such as salient object detection. The nested U-structure of U²-Net enables the model to capture both local and global contextual information at multiple scales, making it particularly well-suited for segmenting intricate tissue structures. Tissue segmentation often requires precise boundary detection and the ability to discern between closely packed or overlapping regions, challenges that U²-Net's deep architecture can effectively address. Additionally, the use of Residual U-blocks (RSU) in U²-Net allows for more efficient learning of hierarchical features, ensuring that both fine details and broader spatial patterns are captured. These characteristics make U²-Net an ideal choice for accurate and robust tissue segmentation, especially in high-resolution images, where traditional models may struggle to maintain performance. We incorporated the implementation and pre-trained model[24] into our built-in tissue segmentation module. Specifically, the model input is an image represented by a matrix $X \in \mathbb{R}^{H \times W \times C}$, where H is the height, W is the width, and C is the number of channels (e.g., RGB). The model output is a matrix $O \in \mathbb{R}^{H \times W \times (C+1)}$, which includes an additional channel to indicate the probability that a pixel belongs to tissue or not. For example, when $c = 3$, i.e., the input is an RGB image, the output becomes an RGBA image, with the last channel representing the transparency value, highlighting the tissue regions.

Vispro module 4: segregating disconnected tissue regions

We identified the connected components within the binary mask using a labeling algorithm from the OpenCV library. This algorithm performs a brute-force search to examine whether neighboring pixels are connected. Specifically, it iterates through each pixel in the binary mask and checks the 4-connectivity (horizontal and vertical neighbors) or 8-connectivity (including diagonal neighbors) to determine if adjacent pixels belong to the same component. Once a group of connected pixels is found, the algorithm assigns a unique integer label to that component. The labeling function outputs a labeled mask, where each connected component is represented by a different integer value.

In our implementation, we added an additional step to remove small, isolated components, which are often caused by staining errors. These small areas are filtered out based on a pre-set area threshold parameter. Any component whose area falls below this threshold is removed from the labeled mask. This refined labeled mask is then used for further analysis and processing of individual tissue regions, allowing for more accurate segmentation and tissue region identification.

Competing methods: 10x standard processing pipeline

In the 10x pipeline, fiducial markers are processed using their proprietary software, such as Space Ranger and Loupe Browser, which generate a layered image. This layered image contains red circles plotted on top of the original H&E-stained tissue image to indicate the fiducial marker areas. These software tools automatically detect and mark the fiducial regions, providing a visual reference for their locations on the tissue.

To integrate this into our workflow for evaluation, we first perform an RGB value thresholding operation on the layered image to isolate the red color component, which represents the fiducial markers. This step extracts the red circles from the image by identifying pixels that fall within a predefined range of red color intensities. Once the red components are isolated, we apply a circle detection algorithm, such as the Hough Circle Transform, to accurately detect the circular fiducial markers. This algorithm scans the image for circular shapes by identifying edge pixels and fitting circles to the detected contours. After detecting the circles, we fill in each marker area to create a binary mask. In this mask, the pixel values are set to 1 within the fiducial marker regions and 0 elsewhere, effectively mimicking the format used in our pipeline. This binary mask is then used for further evaluation, allowing us to directly compare the performance of the 10x pipeline with our own method.

Evaluations

Identifying fiducial markers

To evaluate the accuracy of fiducial marker segmentation, we first generate the ground truth mask from our annotated circle masks, where pixels with a value of 1 indicate the marker areas and pixels with a value of 0 indicate the non-marker areas. We then compute the pixel-level Intersection over Union (IoU) for both our detection results and the results generated by the 10x pipeline, comparing them to the ground truth.

The IoU is a metric that measures the overlap between the predicted mask and the ground truth mask at the pixel level. It is defined as the ratio of the number of correctly predicted pixels (i.e., the intersection of the predicted and ground truth masks) to the total number of pixels in the union of the predicted and ground truth masks:

$$\text{IoU} = \frac{\sum_{i,j} (P_{i,j} \wedge G_{i,j})}{\sum_{i,j} (P_{i,j} \vee G_{i,j})} \tag{7}$$

where $P_{i,j}$ represents the predicted mask at pixel location (i, j), $G_{i,j}$ represents the ground truth mask at pixel location (i, j), $\wedge$ denotes the logical AND operation (i.e., the number of pixels where both $P_{i,j} = 1$ and $G_{i,j} = 1$, representing the intersection), $\vee$ denotes the logical OR operation (i.e., the number of pixels where either $P_{i,j} = 1$ or $G_{i,j} = 1$, representing the union).

A higher IoU value indicates a higher segmentation accuracy, as it shows that the predicted mask aligns more closely with the ground truth mask.

Identifying tissue regions

To evaluate the accuracy of tissue region segmentation, we manually annotated the tissue regions on 30 tissue slices, spanning a range of difficulty from hard to easy samples. We compared three sets of images: the restored images from our method, the original images from the 10x pipeline, and the restored images generated from the 10x pipeline's masks. For each image, we computed the IoU with the corresponding ground truth.

Segregating disconnected tissue regions

To evaluate the accuracy of tissue segregation, we manually annotated the disconnected tissue regions on XX slices, each containing more than one distinct tissue area. We then compared the number of segmented tissue regions produced by our method and by the 10x pipeline, and calculated the L1-distance between the ground truth and the model output to measure the accuracy. The L1-distance is defined as:

$$\text{L1-distance} = \left| T^{\text{pred}} - T^{\text{true}} \right| \tag{8}$$

where T^{pred} represents the number of segmented tissue regions predicted by the model for the slice, and T^{true} represents the number of tissue regions in the ground truth. A lower L1-distance indicates higher tissue segregation accuracy, as it reflects a smaller discrepancy between the predicted and actual number of tissue regions.

Cell segmentation

To evaluate cell segmentation performance, we used StarDist to detect and count the number of segmented cells in areas outside the tissue. Since we have manually annotated the tissue regions, no cells should be detected in the out-of-tissue areas. Thus, a lower number of segmented cells in these regions indicates better segmentation performance.

Image registration

To evaluate the accuracy of image registration on both same-modal and cross-modal image registration. For the same-modal setting, we visually selected 20 pairs of Visium images from our collected dataset. These images are all H&E stained, with neighboring images sharing similar textures. For the cross-modal setting, we collected 13 groups from the Visium CytAssist dataset. The Visium CytAssist dataset involves overlaying images of standard histological glass slides and those processed into Visium slides. The two images are structurally the same but display some differences in textures, making them suitable for cross-modal registration tests.

We employed the tool of bUnwarpJ[18] for image registration. Two metrics are utilized to evaluate the results which are commonly employed to evaluate registration performance. The first one is the Structural Similarity Index (SSIM), which compares local patterns of pixel intensities that have been normalized for luminance and contrast. This metric is particularly sensitive to pixel misalignment within same-model image registration setting. It is defined as:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{9}$$

where μ_x, μ_y, σ_x, σ_y, and σ_{xy} are the local means, standard deviations, and cross-covariance for images x and y.

The second metric is Mutual Information (MI), a measure of the mutual dependence between the two images, which proves to be especially useful for cross-modal image registration evaluation. It is defined as:

$$\mathrm{MI}(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right) \tag{10}$$

where $p(x,y)$ is the joint probability distribution function, and $p(x)$, $p(y)$ are the marginal probability distributions of X and Y.

Acknowledgments

The study was supported by the National Institutes of Health under Award Number 1U54AG075936-01.

Author contributions

H.M., A.Z. and Z.J. conceived the study. H.M. and Y.Q. collected the training data. H.M. conducted the analysis. H.M. and Z.J. wrote the manuscript.

Competing interests

All authors declare no competing interests.

Data availability

All data generated or analysed during this study are included in this published article and its supplementary information files.

Code availability

Vispro can be freely accessible at https://github.com/HuifangZJU/Vispro. The Github repository includes detailed instructions for usage and any required dependencies.

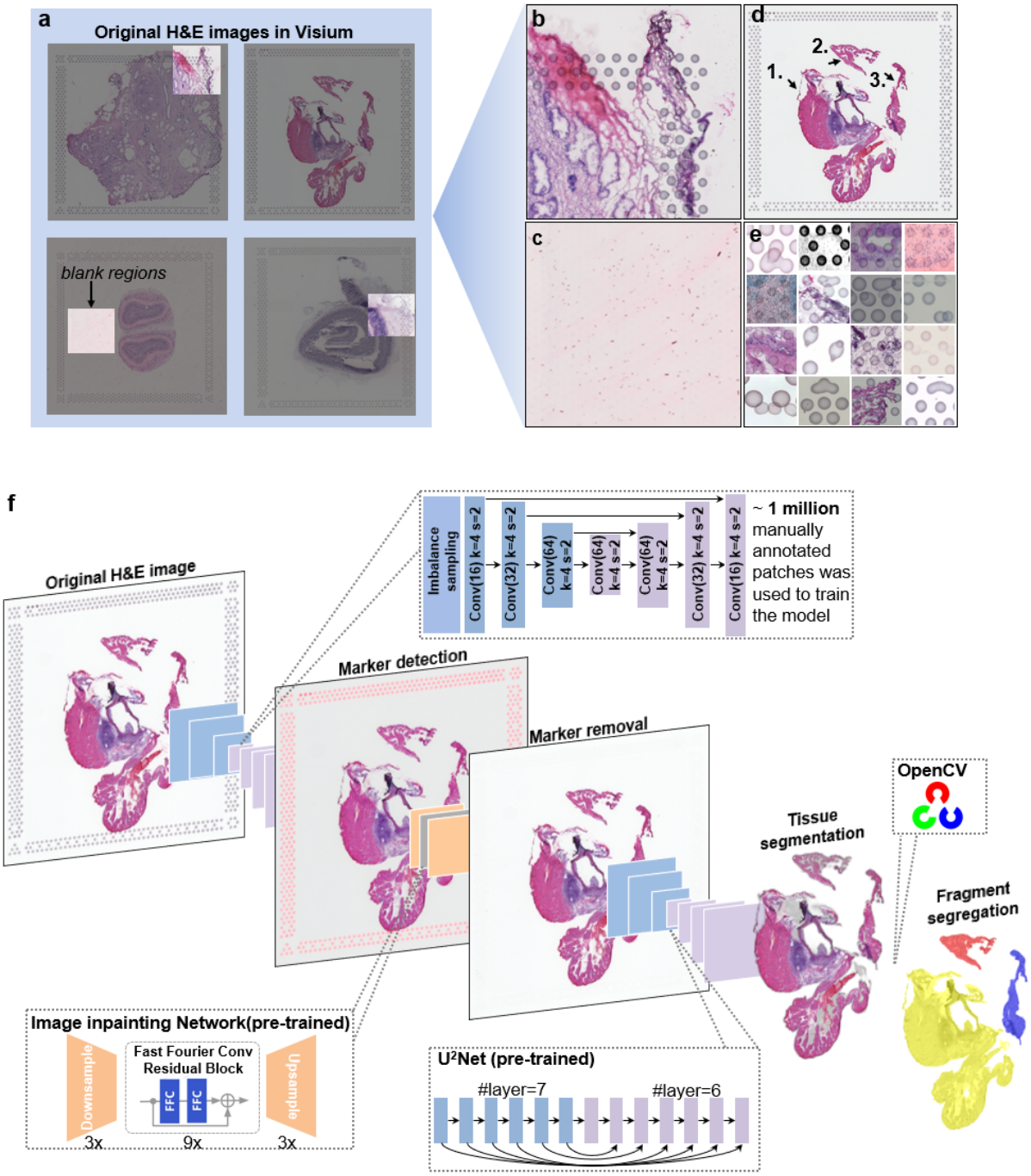

Figure 1. Overview of the Vispro analysis pipeline. a. Original H&E-stained tissue images captured in the Visium platform, highlighting regions of interest. b. A close-up of tissue boundaries with marker overlays. c. Background noise and marker artifacts are shown in the blank regions on the Visium image. d. Different tissue subregions marked as 1, 2, and 3. e. Examples of deformation in fiducial markers. f. The complete methodological workflow, including marker detection, marker removal, tissue segmentation and tissue segregation.

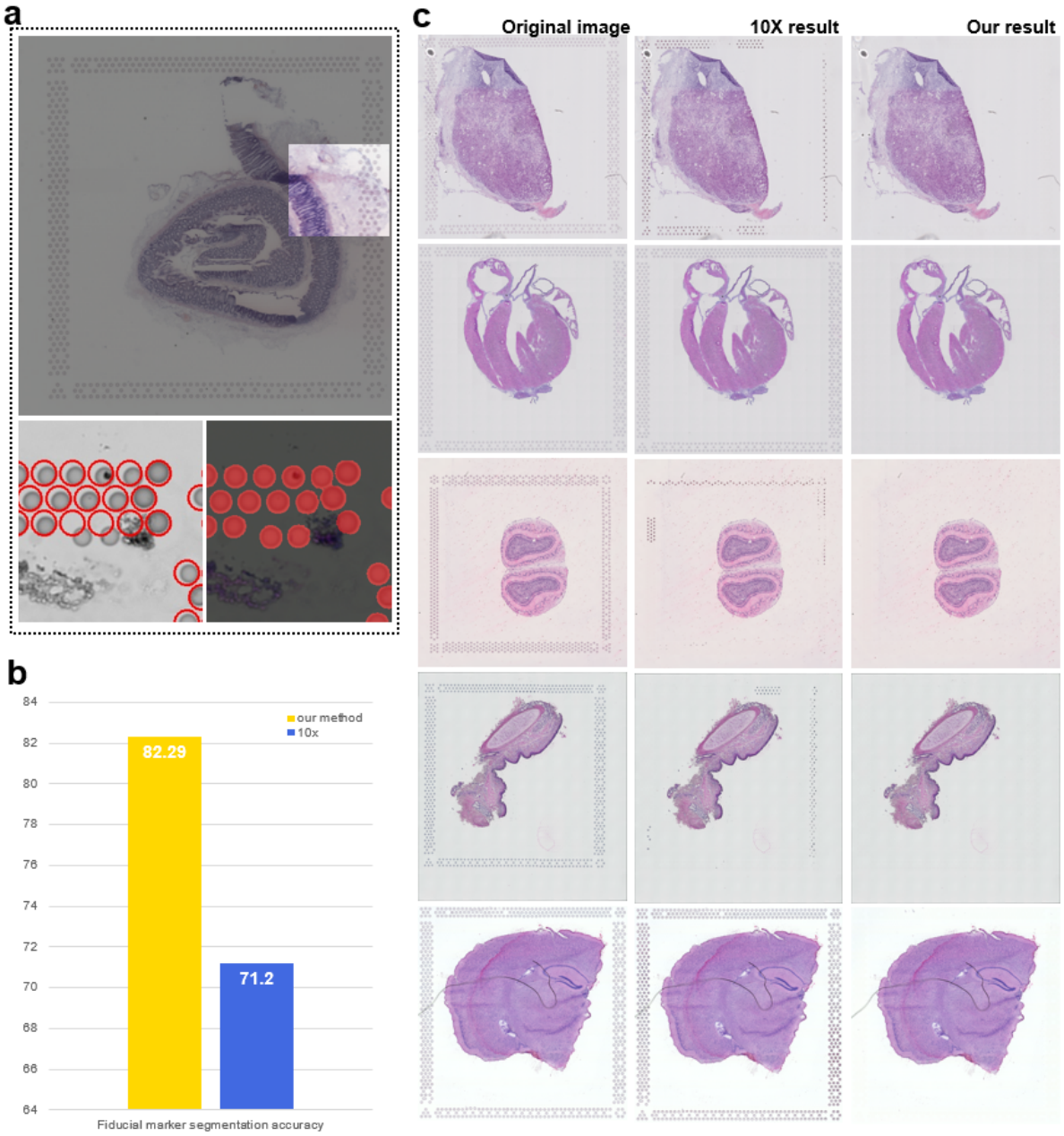

Figure 2. Improved fiducial marker segmentation in spatial transcriptomics data. a. Tissue slice with detected fiducial markers (red circles), including zoomed-in views. b. Bar chart comparing segmentation accuracy: our method (82.29%) vs. 10x (71.2%). c. Visual comparison of segmentation results across different tissue samples: Original image, 10x mask, and our mask.

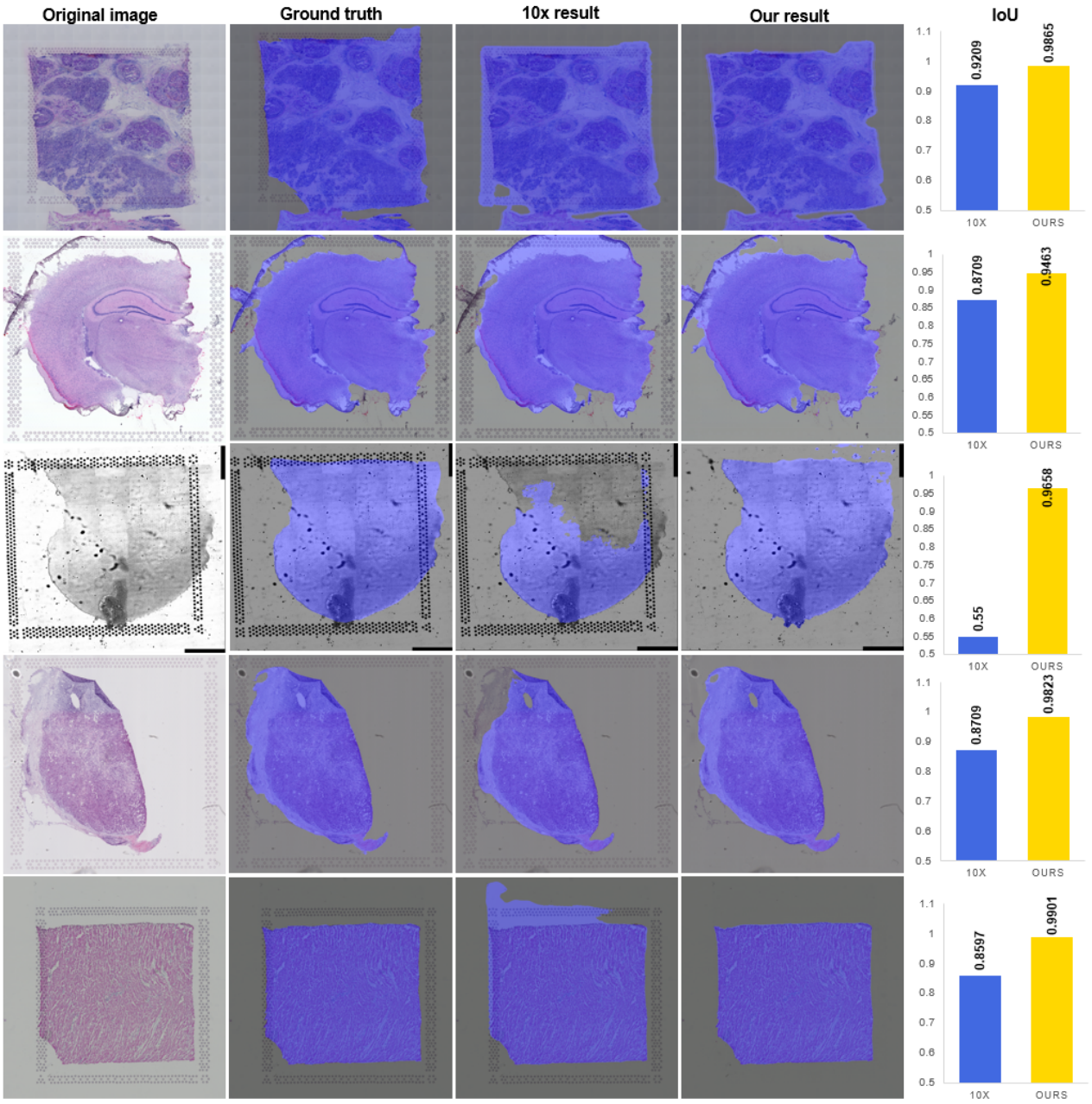

Figure 3. Tissue segmentation results comparing ground truth, 10x result, and our method. Rows show different tissue samples with detection overlays. Bar graphs on the right compare IoU scores between 10x (blue) and our method (yellow), demonstrating improved performance across samples

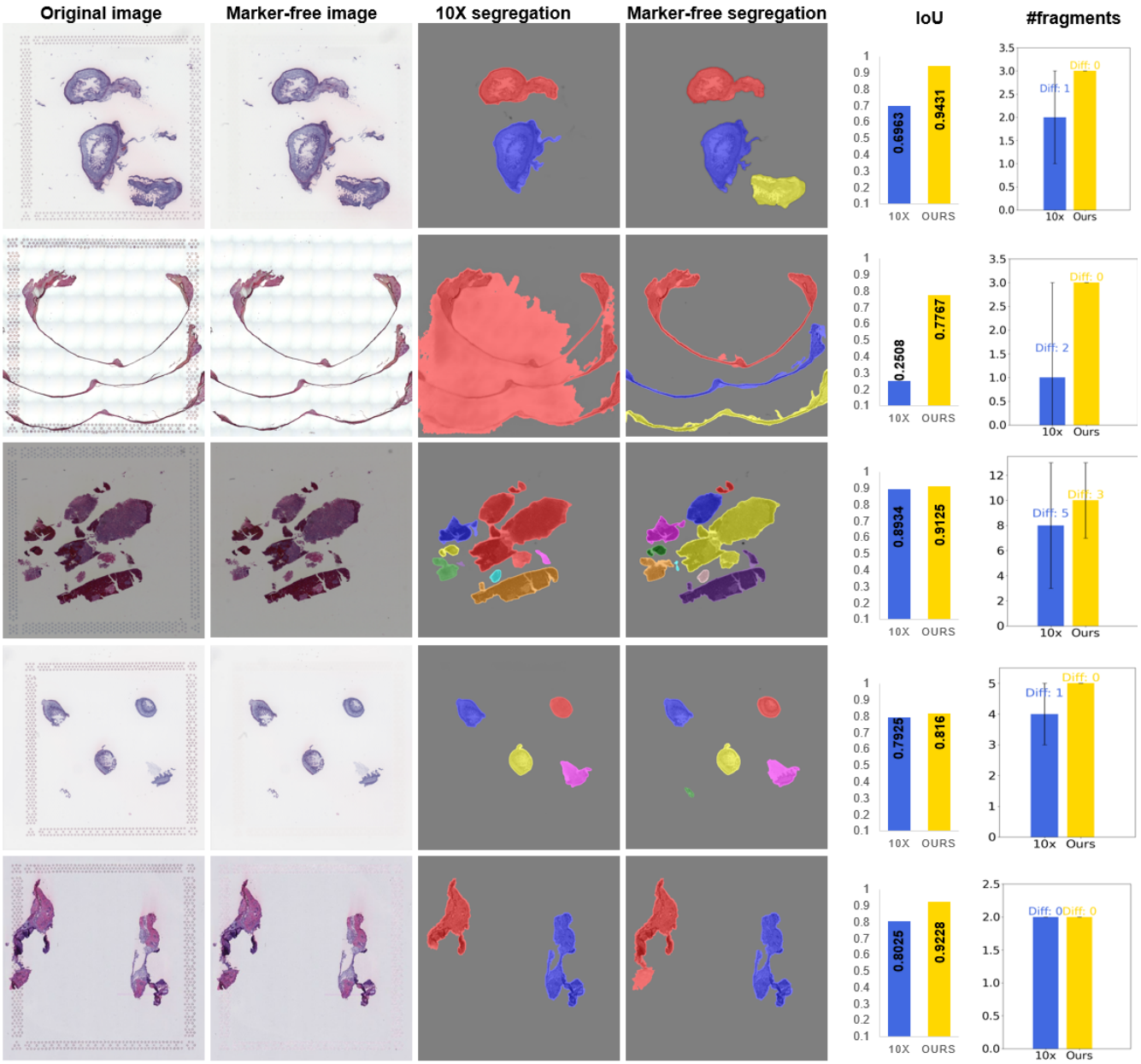

Figure 4. Tissue segregation result comparing 10X-based segregation with our method. The original images (first column) are shown alongside marker-free images (second column), followed by the 10X segregation results (third column) and our marker-free segregation method (fourth column). The last two columns display the intersection-over-union (IoU) scores and the number of detected fragments for both methods. Our approach demonstrates improved tissue segregation with fewer fragments, as evidenced by consistently higher IoU values and closer-to-true fragment counts across different tissue samples.

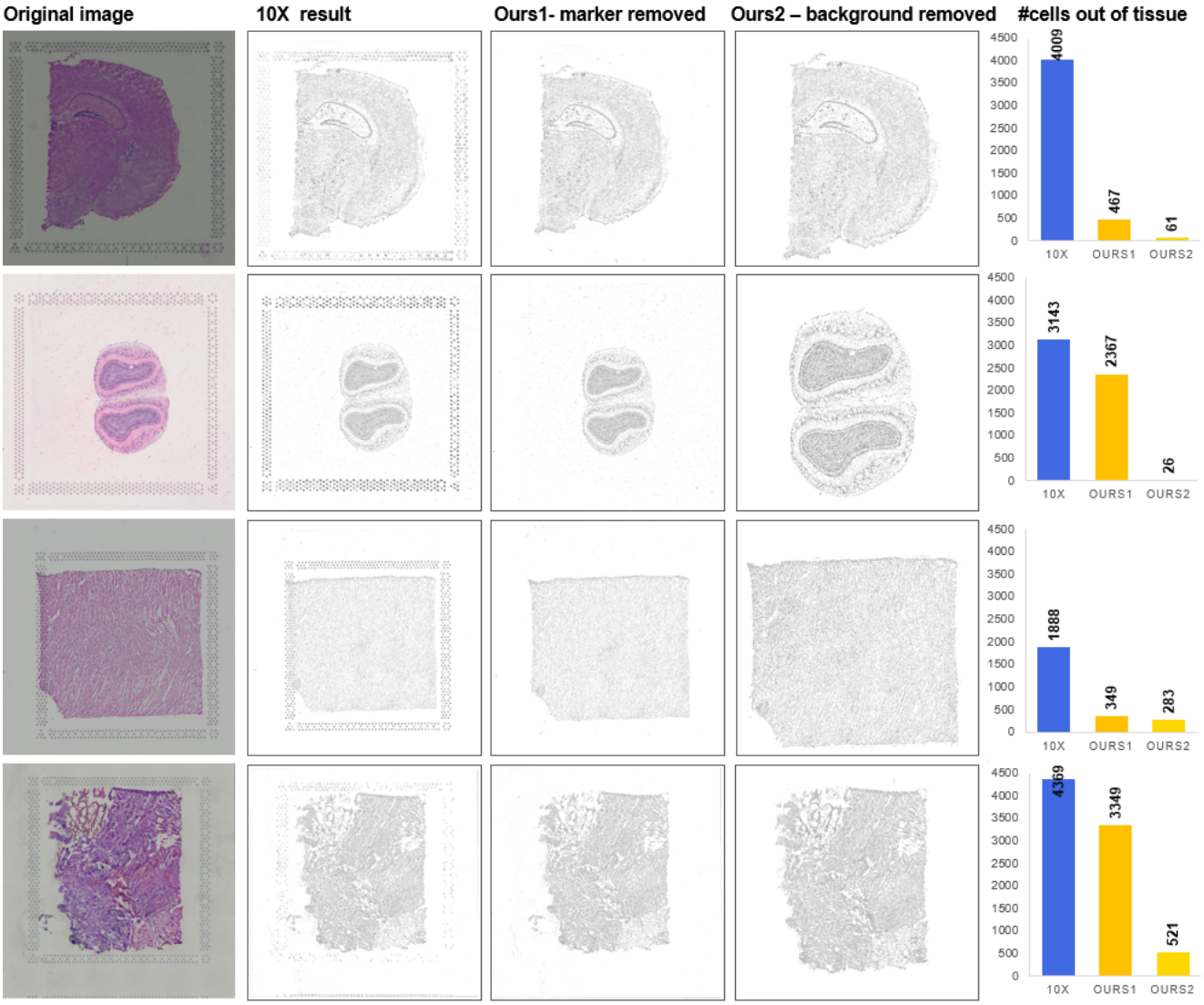

Figure 5. Comparison of out-of-tissue cell segmentation result. The original images (first column) are shown alongside the results of 10X-based cell segmentation (second column), our marker-removed approach (third column), and our background-removed approach (fourth column). The final column presents the number of cells segmented outside the tissue regions, highlighting the performance improvements of our methods in reducing spurious cell segmentations. Our approach (Ours2) effectively minimizes out-of-tissue cell detections compared to the 10X method, showcasing enhanced precision in tissue-specific cell localization.

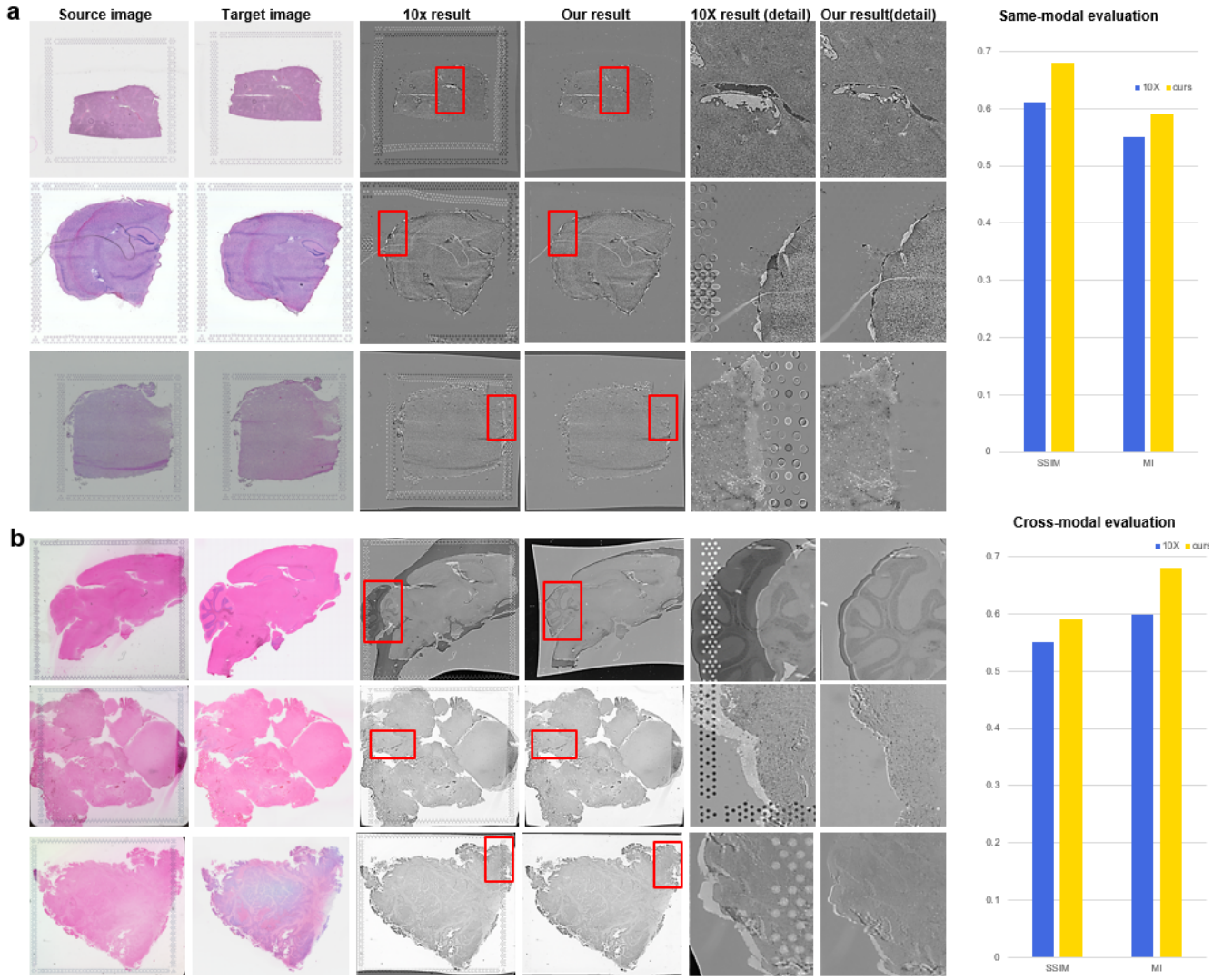

Figure 6. Comparison of registration and detail recovery in same- and cross-modal evaluations. a. Registration results using the 10X method and our approach for same-modal image registration. The source and target images are shown alongside the registration results. Red boxes highlight regions of interest, with zoomed-in views comparing the 10X result and our method for detail recovery. The bar chart on the right compares structural similarity index (SSIM) and mutual information (MI) metrics, showing improved performance of our method. b. Cross-modal image registration comparison using 10X and our method. The results and detailed regions (red boxes) are displayed, with the right bar chart showing evaluation metrics, indicating superior cross-modal alignment by our approach.

References

1. Pang, M., Su, K. & Li, M. Leveraging information in spatial transcriptomics to predict super-resolution gene expression from histology images in tumors. *BioRxiv* 2021–11 (2021).
2. Zeng, Y. *et al.* Spatial transcriptomics prediction from histology jointly through transformer and graph neural networks. *Briefings Bioinforma.* **23**, bbac297 (2022).
3. Jia, Y., Liu, J., Chen, L., Zhao, T. & Wang, Y. Thitogene: a deep learning method for predicting spatial transcriptomics from histological images. *Briefings Bioinforma.* **25**, bbad464 (2024).
4. Xie, R. *et al.* Spatially resolved gene expression prediction from histology images via bi-modal contrastive learning. *Adv. Neural Inf. Process. Syst.* **36** (2024).
5. Zeng, Y. *et al.* Identifying spatial domain by adapting transcriptomics with histology through contrastive learning. *Briefings Bioinforma.* **24**, bbad048 (2023).
6. Mejia, G., Cárdenas, P., Ruiz, D., Castillo, A. & Arbeláez, P. Sepal: Spatial gene expression prediction from local graphs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2294–2303 (2023).
7. Xiao, X., Kong, Y., Li, R., Wang, Z. & Lu, H. Transformer with convolution and graph-node co-embedding: an accurate and interpretable vision backbone for predicting gene expressions from local histopathological image. *Med. Image Analysis* **91**, 103040 (2024).
8. Fu, X. *et al.* Spatial gene expression at single-cell resolution from histology using deep learning with ghist. *bioRxiv* 2024–07 (2024).
9. Hu, J. *et al.* Deciphering tumor ecosystems at super resolution from spatial transcriptomics with tesla. *Cell systems* **14**, 404–417 (2023).
10. Hu, J. *et al.* Spagcn: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nat. methods* **18**, 1342–1351 (2021).
11. Pham, D. *et al.* stlearn: integrating spatial location, tissue morphology and gene expression to find cell types, cell-cell interactions and spatial trajectories within undissociated tissues. *BioRxiv* 2020–05 (2020).
12. Daly, A. C., Geras, K. J. & Bonneau, R. A convolutional neural network for common coordinate registration of high-resolution histology images. *Bioinformatics* **37**, 4216–4226 (2021).
13. 10x Genomics. Space ranger (version 2.0.0). https://support.10xgenomics.com/spatial-gene-expression/software/pipelines/latest/what-is-space-ranger (2023). Accessed: October 5, 2024.
14. 10x Genomics. Loupe browser (version 7.0.0). https://support.10xgenomics.com/spatial-gene-expression/software/visualization/latest/what-is-loupe-browser (2023). Accessed: October 5, 2024.
15. Wang, Y. *et al.* A systematic evaluation of computational methods for cell segmentation. *Briefings Bioinforma.* **25**, bbae407 (2024).
16. Schmidt, U., Weigert, M., Broaddus, C. & Myers, G. 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, 2018).
17. Pardo, B. *et al.* spatiallibd: an r/bioconductor package to visualize spatially-resolved transcriptomics data. *BMC genomics* **23**, 434 (2022).
18. Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nat. methods* **9**, 676–682 (2012).
19. Duda, R. O. & Hart, P. E. Use of the hough transformation to detect lines and curves in pictures. *Commun. ACM* **15**, 11–15 (1972).
20. Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988 (2017).
21. Suvorov, R. *et al.* Resolution-robust large mask inpainting with fourier convolutions. *arXiv preprint arXiv:2109.07161* (2021).
22. Iglovikov, V. I. & Seferbekov, A. lama: Resolution-robust large mask inpainting with fourier convolutions. https://github.com/advimman/lama (2021). Accessed: October 5, 2024.

23. Qin, X. *et al.* U2-net: Going deeper with nested u-structure for salient object detection. *Pattern recognition* **106**, 107404 (2020).

24. Manira, N. E. Background remover: A tool for removing background from images. https://github.com/nadermx/backgroundremover (2023). Accessed: October 5, 2024.