MEATRD: Multimodal Anomalous Tissue Region Detection Enhanced with Spatial Transcriptomics

Kaichen Xu[1*], **Qilong Wu** [1*], **Yan Lu** [1], **Yinan Zheng** [1], **Wenlin Li** [1], **Xingjie Tang** [1], **Jun Wang** [2], **Xiaobo Sun** [1†]

[1] School of Statistics and Mathematics, Zhongnan University of Economics and Law
[2] iWudao Tech
{kaichenxu, qilongwu, yanlu, yinanzheng, wenlinli, xingjietang}@stu.zuel.edu.cn, xsun28@gmail.com, jwang@iwudao.tech

Abstract

The detection of anomalous tissue regions (ATRs) within affected tissues is crucial in clinical diagnosis and pathological studies. Conventional automated ATR detection methods, primarily based on histology images alone, falter in cases where ATRs and normal tissues have subtle visual differences. The recent spatial transcriptomics (ST) technology profiles gene expressions across tissue regions, offering a molecular perspective for detecting ATRs. However, there is a dearth of ATR detection methods that effectively harness complementary information from both histology images and ST. To address this gap, we propose MEATRD, a novel ATR detection method that integrates histology image and ST data. MEATRD is trained to reconstruct image patches and gene expression profiles of normal tissue spots (inliers) from their multimodal embeddings, followed by learning a one-class classification AD model based on latent multimodal reconstruction errors. This strategy harmonizes the strengths of reconstruction-based and one-class classification approaches. At the heart of MEATRD is an innovative masked graph dual-attention transformer (MGDAT) network, which not only facilitates cross-modality and cross-node information sharing but also addresses the model over-generalization issue commonly seen in reconstruction-based AD methods. Additionally, we demonstrate that modality-specific, task-relevant information is collated and condensed in multimodal bottleneck encoding generated in MGDAT, marking the first theoretical analysis of the informational properties of multimodal bottleneck encoding. Extensive evaluations across eight real ST datasets reveal MEATRD's superior performance in ATR detection, surpassing various state-of-the-art AD methods. Remarkably, MEATRD also proves adept at discerning ATRs that only show slight visual deviations from normal tissues.

Code — https://github.com/wqlzuel/MEATRD
Extended version — https://arxiv.org/abs/2412.10659

Introduction

Detecting anomalous tissue regions (ATR) within tissues from affected individuals is essential in clinical diagnostics, pathological studies, and targeted therapies (Srinidhi, Ciga, and Martel 2021). Traditionally, automated ATR detection, which typically applies computer vision techniques to histology images, e.g., whole-slide images (WSI) stained with hematoxylin and eosin (H&E) (Zingman et al. 2023), is a specialized task of image anomaly detection (AD). However, histology images, unlike natural images (e.g., those in ImageNet dataset) (Bergmann et al. 2019), present unique challenges for AD due to their inherent high complexity (Zehnder et al. 2022), subtle differences between ATRs and normal tissues (Shenkar and Wolf 2022), the diverse manifestations of ATRs (Komura and Ishikawa 2018), and variability in staining quality (Zingman et al. 2023). The complexities demand supplementary information to visual cues for accurate ATR detection.

Spatial transcriptomics (ST) meets this need by providing spatial gene expression data. By far, a total of 1033 publicly available human ST datasets that span 56 diseases and 35 tissues, providing a rich resource for investigating ATRs at the molecular level (Wang et al. 2024). A typical ST dataset is structured as a matrix $\mathbf{X} \in \mathbb{R}^{N \times G}$, where $\mathbf{X}_{i,j}$ denotes the expression read counts of the j-th gene mapped to i-th tissue spot. As illustrated in Figure 1, these spots, ranging in size from 10 to 200 μm as per sequencing technology, are spatially arranged in arrays to cover the entire tissue slice (Hu et al. 2023), thereby characterizing gene expression profile across the tissue. This molecular-level data, especially in cases where ATRs are visually similar to normal tissues, can significantly aid in their detection (Hu et al. 2021). However, due to limitations inherent to sequencing technology, ST data suffer from severe noise and substantial missing values in gene expression measurements (Wang et al. 2022), leading to compromised precision in demarcating tissue regions (Wang, Maletic-Savatic, and Liu 2022). Integration of histology images with ST data presents a promising solution to these challenges. As illustrated in our toy example in Figure 1, the blank spots in the ST dataset's spatial map, which represent tumor core locations with missing gene expression data, are visually identifiable in the accompanying histology image. Conversely, the tumor edge region, which may not be easily distinguishable from normal tissues visually, is detectable in the ST data. Therefore, the information from the two modalities can complement each other, greatly enhancing the precision of ATR detection. Fortunately, ST technologies like 10x Visium (Moses and Pachter 2022) provide accom-

[*]These authors contributed equally.
[†]Corresponding author.

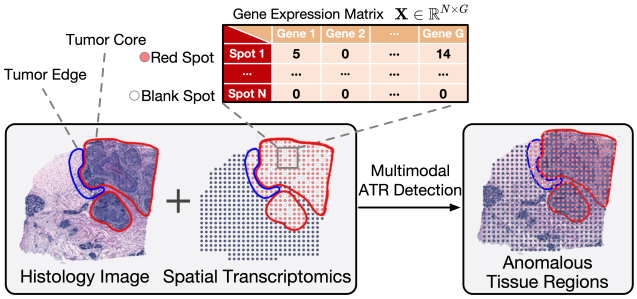

Figure 1: Detecting ATRs with histology images and ST data. ATRs include both tumor core and edge regions, as delineated by red and blue outlines in the histology image, respectively. The tumor edge region visually resembles the adjacent normal tissues. In the spatial map of the ST dataset, the ATRs encompass both red and blank spots, with blank spots indicating locations of missing gene expression data.

panying histology images, allowing concurrent analysis of visual and genetic information for ATR detection.

Given the rarity and unpredictable heterogeneity of anomalies, AD in images is often framed as an unsupervised learning problem, where anomalies are not known a priori. Models are trained exclusively on reference datasets comprising inliers to understand "normality" at training time and identify deviations from this norm as anomalies at inference time (Liu et al. 2023; Bergmann et al. 2019). Contemporary image AD methods, which use deep learning to learn initial representations of normal images (Shvetsova et al. 2021; Liu et al. 2023), often involve an encoder pre-trained on large natural image datasets (Deng and Li 2022; Roth et al. 2022). These representations are then used to either model the inlier distribution in latent space, as seen in one-class classification methods (Ruff et al. 2018), or to reconstruct inliers in reconstruction-based methods (Schlegl et al. 2019). Instances in the target dataset, which exhibit low probability in the inlier distribution or larger-than-expected reconstruction errors are deemed anomalous.

Despite successes of these methods in areas such as manufacturing defect inspection, financial fraud detection, etc (Sohn et al. 2020), the unique challenges posed by ATR detection require more specialized methods (Riasatian et al. 2021; Tschuchnig and Gadermayr 2022). To meet this need, adaptions made to image AD methods focus on representation learning and anomaly discrimination techniques. For example, image encoders pre-trained on natural images are replaced with those tailored for histology images, such as U-Net (Zehnder et al. 2022), DenseNet (Riasatian et al. 2021), and s2-AnoGAN (Pocevičiūtė, Eilertsen, and Lundström 2021). In addition, anomaly scoring is adapted to use perceptual loss instead of pixel-wise reconstruction errors commonly used for natural images (Shvetsova et al. 2021; Zehnder et al. 2022). However, these methods may struggle when ATRs visually resemble normal tissues (Bejnordi et al. 2017). In contrast, ST differentiates tissue regions at the gene expression level (Hu et al. 2021; Dong and Zhang 2022), providing a remedy for ATR detection involving such complexities. Currently, Spatial-ID (Shen et al. 2022) is the only method that uses ST for ATR detection, employing a DNN classifier which assigns spots in the ST dataset to known regions while determining those with uncertain assignments as anomalies. However, this classification-based approach can induce high false positive rates, as uncertainties in assignment could stem from similarities among normal tissues rather than the presence of ATRs (Li et al. 2022). Its sole reliance on ST data also makes it vulnerable to noise and dropouts in gene expression measurements, even for detecting visually recognizable ATRs.

In this study, we propose **M**ultimodality **E**nhanced **A**nomalous **T**issue **R**egion **D**etection (**MEATRD**), the first method that integrates histology images and ST data for enhanced ATR detection. MEATRD conceptualizes tissue spots as nodes within an attributed graph, leveraging a reconstruction-based graph model for inlier nodes reconstruction from dual perspectives of image and gene expression. During inference, the discrepancies between reconstruction errors of inliers (i.e., normal tissues) and anomalies (i.e., ATRs) can be exploited by a discriminative model for accurate ATR detection. As shown in Figure 2, MEATRD involves three stages. **Stage I** focuses on extracting visual features of histology images. The histology image is segmented into a patch centered around each spot, which are processed into imagery embeddings. **Stage II** aims to reconstruct the gene expression profiles and image patches of each spot from their fused embeddings, obtained using our innovative masked graph dual-attention transformer (MGDAT) network. MGDAT allows concurrent cross-node and cross-modal attention calculations, promoting efficient cross-modality information sharing and incorporation of spatial relationships among spots. Additionally, to counter potential model over-generalization[1], we employ the node-feature masking strategy, which forces the model to rely more on the surrounding context and cross-modal information. **Stage III** focuses on acquiring a one-class classification model to identify anomalies. Unlike existing one-class classification AD methods that use instance deep embeddings and are prone to reference-target domain shifts (Ouardini et al. 2019), our model pioneers in using domain shift-robust latent multimodal reconstruction losses (Donahue, Krähenbühl, and Darrell 2016; Schlegl et al. 2019) for more reliable anomaly detection. By collapsing inliers' reconstruction losses into a compact hypersphere, our model increases the reconstruction error discrepancy between inliers and anomalies, thereby further mitigating model over-generalization. In summary, our main contributions include:

- We propose MEATRD, a pioneer multimodal method that integrates spatial transcriptomics with histology images for enhanced ATR detection.
- MEATRD simultaneously addresses the over-generalization in reconstruction-based AD methods and the domain shift issue in one-class classification, leading to significant performance improvement.

[1]A common pitfall of reconstruction-based methods where anomalies might yield low reconstruction errors (Liu et al. 2023; Ristea et al. 2022).

- We design an MGDAT network as the core component of MEATRD to facilitate cross-modality and cross-node information exchange while ameliorating model over-generalization. We also demonstrate the theoretical foundation for this information exchange, which is grounded in MGDAT's ability to generate inclusive and condensed encoding of modality-specific, task-relevant information (supplementary material D).
- Extensive benchmarks on eight breast cancer ST datasets demonstrate MEATRD's superiority over nine state-of-the-art (SOTA) AD methods in accurately detecting ATRs, including those with subtle visual deviations from surrounding normal tissues.

Preliminary

Definition D.1. ***ST Dataset and Associated Histology Image.*** *Let $\boldsymbol{X} \in \mathbb{R}^{N \times G}$ be a ST dataset, where N is the number of tissue spots and G is the number of genes. S_N and S_G denote the set of spots and genes, respectively. $\boldsymbol{X}_{i,j}$ represents the the read counts of gene j at spot i, and $\boldsymbol{x}_i \in \mathbb{R}^G$ represents the gene expression profile at spot i. Let $\boldsymbol{P} \in \mathbb{R}^{H \times W \times C}$ be the associated histology image, where H, W, and C are the height, width, and number of channels, respectively.*

Definition D.2. ***Graph Representation of ST Dataset and Histology Image.*** *For a given ST dataset $\boldsymbol{X}$, the associated histology image $\boldsymbol{P}$ is segmented into N patches, where $\boldsymbol{P}_i \in \mathbb{R}^{h \times w \times C}$ denotes the patch centered around spot $i \in S_N$, with height h and width w. Then spots are modeled as nodes on an unweighted, attributed graph $G(S_N, A, \mathcal{Z})$, where $A \in \{0,1\}^{N \times N}$ is the adjacency matrix, and $\mathcal{Z} \coloneqq [\mathcal{Z}_{image} || \mathcal{Z}_{gene}]$ is the node feature matrix. $\mathcal{Z}_{img} \in \mathbb{R}^{N \times D}$ and $\mathcal{Z}_{gene} \in \mathbb{R}^{N \times D}$ are embeddings of image patches and gene expression profiles of spots. $A(i,j) = 1$ if node $j \in n(i)$, where $n(i)$ is the set of k-nearest neighbors of node i, and $A(i,j) = 0$ otherwise. k is typically set to be six due to the hexagonal arrangement of spots (Xu et al. 2024).*

Definition D.3. ***Problem Definition.*** *Let $\mathcal{X}$ and $\mathcal{P}$ denote the target ST dataset and associated histology image, respectively. Similarly, let $\boldsymbol{X}$ and $\boldsymbol{P}$ denote the reference ST dataset and associated histology image, respectively. We define $y_i \in \{0,1\}$ as the label for spot i, where $y_i = 1$ indicates an anomalous spot, and $y_i = 0$ otherwise. Note, $y_i = 0, \forall i \in \boldsymbol{X}$; $y_j \in \{0,1\}, \forall j \in \mathcal{X}$. The task of ATR detection is defined as identifying the subset of anomalous spots within the target dataset: $\mathbb{S} = \{\S_i | y_{\S_i} = 1, \forall \S_i \in \mathcal{X}\}$, using a model trained exclusively on $\boldsymbol{X}$ and $\boldsymbol{P}$.* [2]

Method

As shown in Figure 2, the workflow of MEATRD includes three stages: Stage I extract visual features from histology image patch of each spot; Stage II focuses on the learning of reconstructions of image patches and gene expression profiles using multimodal embeddings generated by a MGADT network; Stage III entails the training of an anomaly discriminator based on latent multimodal reconstruction errors.

[2]Related work is in supplementary material A due to space limitation.

Extracting Visual Features of Histology Image Patches (Stage I)

Initially, whole slide images are segmented into 32x32 patches centered around each spot in the ST dataset (Zong et al. 2022). The visual manifolds of these image patches are obtained using a Mobile-Unet, with an encoder consisting of downsampling convolutional layers and inverted residual blocks. Its decoder comprises upsampling deconvolutional layers and inverted residual blocks, connected to the encoder via shortcut connections.

This design not only inherits the merits of U-Net in extracting visual features from histology images but also boosts computational efficiency by reducing the model's parameters. Given a histology image patch $\boldsymbol{P}_i$ for spot $i \in S_N$, the Mobile-Unet is pretrained to reconstruct it as $\widehat{\boldsymbol{P}}_i$, with a pretraining loss that is a mix of a perceptual loss, based on the Structural Similarity Index (SSIM), and an $L1$ reconstruction loss:

$$\widehat{\boldsymbol{P}}_i \coloneqq D_1(E_1(\boldsymbol{P}_i)), \quad \boldsymbol{z}_i \in \mathbb{R}^D \coloneqq E_1(\boldsymbol{P}_i) \tag{1}$$

$$\mathcal{L}_{perc} = -\text{SSIM}(\boldsymbol{P}_i, \widehat{\boldsymbol{P}}_i), \mathcal{L}_1 = ||\boldsymbol{P}_i - \widehat{\boldsymbol{P}}_i||_1 \tag{2}$$

$$\text{SSIM}(\boldsymbol{X}, \boldsymbol{Y}) = \frac{(2\mu_{\boldsymbol{X}}\mu_{\boldsymbol{Y}} + C_1)(2\sigma_{\boldsymbol{X},\boldsymbol{Y}} + C_2)}{(\mu_{\boldsymbol{X}}^2 + \mu_{\boldsymbol{Y}}^2 + C_1)(\sigma_{\boldsymbol{X}}^2 + \sigma_{\boldsymbol{Y}}^2 + C_2)} \tag{3}$$

$$\mathcal{L}_{pretrain} = \mathcal{L}_{perc} + \mathcal{L}_1. \tag{4}$$

where μ_* and σ_*^2 are the average intensity and variance of $* \in \{\boldsymbol{X}, \boldsymbol{Y}\}$, respectively. C_1 and C_2 represent two constants to stabilize the division with a weak denominator. SSIM and $\mathcal{L}_1$ measure the structural similarities and pixel-by-pixel discrepancies between the original and reconstructed images, respectively. Then, pretraining loss enhances the representation learning of complex histology images by accounting for both contextual integrity, via $\mathcal{L}_{perc}$, and local details via $\mathcal{L}_1$ (Okada and Taniguchi 2021). Following training, E_1 is used to yield image patch embeddings for each spot $i \in S_N$. Finally, unlike complex tissue images, which need to be converted into semantically meaningful representations in the first place, gene data have much clearer semantics. Therefore, MEATRD do not require a pretext representation learning stage for gene data. Rather, we use a two-layer MLP in stage II to rasterize gene data before feeding them into MGDAT blocks, where graph-based gene encoding takes places.

Masked Graph Dual-Attention Transformer Network (Stage II)

To generate information-rich multimodal spot embeddings for reconstruction, we fuse histology image patches and gene expression profiles while incorporating contextual interdependencies among spots to reveal their biological characteristics. This is achieved by modeling spots as nodes in an attributed graph $G(V, A, \mathcal{Z})$, as described in Definition D.2, on top of which node representations are learned using an innovative masked graph attention network, termed MGDAT. This network, comprising a series of MGDAT blocks, allows information sharing across both data modality and graph nodes. Within each MGDAT block, nodes to be reconstructed are

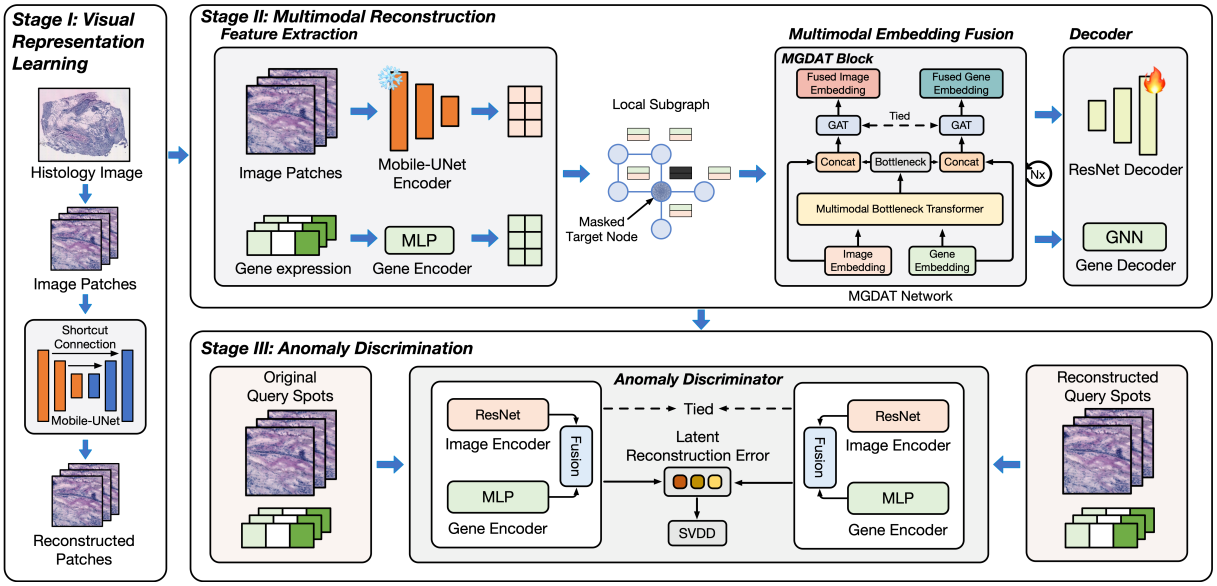

Figure 2: The workflow of MEATRD.

masked before aggregating fused gene and imagery attributes of their neighboring nodes via attention-based mechanism.

Formally speaking, let $G_i(V_i, A_i, \mathcal{Z}_i)$ denote the subgraph of a target node i that covers up to its 3-hop neighbors, where V_i, A_i and $\mathcal{Z}_i$ denote the node set, adjacency matrix, and node attribute matrix of G_i, respectively. We set the number of hops to be 3 as using more hops will result in over-smoothing while fewer hops will significantly limit the information spread in the graph. $\boldsymbol{z}_i \in \mathbb{R}^D$ represents node i's imagery attribute derived from Stage I, and $\boldsymbol{\zeta}_i \in \mathbb{R}^D$ represents node i's gene attribute rasterised from its gene expression vector $\boldsymbol{x}_i$ using a two-layer MLP. $\boldsymbol{z}_i$ and $\boldsymbol{\zeta}_i$ are substituted with learnable mask tokens $\boldsymbol{z}_{[M]} \in \mathbb{R}^D$and $\boldsymbol{\zeta}_{[M]} \in \mathbb{R}^D$.

This target-node-masking serves to prevent self-information leakage of the target node into its own embedding for reconstruction, thus alleviating the potential model over-generalization issue. G_i is processed by the MGDAT network through its series of MGDAT blocks. For the l-th block, $l \in \{0, 1, 2\}$, the inputs are embeddings of the image patches, $\mathcal{Z}^{(l)}_{img,i} \in \mathbb{R}^{|V_i| \times D}$, and the gene expression profiles, $\mathcal{Z}^{(l)}_{gene,i} \in \mathbb{R}^{|V_i| \times D}$, of V_i. The initial embeddings are defined as $\mathcal{Z}^{(0)}_{img,i} := [\boldsymbol{z}_1, .., \boldsymbol{z}_{[M]}, ..\boldsymbol{z}_{V_i}]^\top$ and $\mathcal{Z}^{(0)}_{gene,i} := [\boldsymbol{\zeta}_1, .., \boldsymbol{\zeta}_{[M]}, .., \boldsymbol{\zeta}_{V_i}]^\top$. The l-th MGDAT block yields fused bottleneck embeddings $\mathcal{Z}^{(l)}_{fb,i} \in \mathbb{R}^{|V_i| \times D'}$, $D' \ll D$ as follows:

$$\mathcal{Z}^{(l)}_{fb,i} = \text{Trm}\left([\mathcal{Z}^{(l)}_{img,i} || \mathcal{Z}^{(l)}_{gene,i}]; W^{(l)}_Q, W^{(l)}_K, W^{(l)}_V\right) \quad (5)$$

where Trm denotes Transformer. $W^{(l)}_Q, W^{(l)}_K, W^{(l)}_V \in \mathbb{R}^{2D \times D'}$ are query, key and value parameters, respectively. $\mathcal{Z}^{(l)}_{fb}$ serves as a bottleneck to collate and condense modality-specific, task-relevant information from image and ST data (Nagrani et al. 2021), as theoretically demonstrated in supplementary material D. By concatenating $\mathcal{Z}^{(l)}_{fb}$ with $\mathcal{Z}^{(l)}_{img}$ and $\mathcal{Z}^{(l)}_{gene}$, the two data modalities are bridged, facilitating access to their complementary task-relevant information. Next, multimodal information of l-hop neighbors is aggregated as follows:

$$h^{(l)}_{*,i} = [\mathcal{Z}^{(l)}_{*,i} || \mathcal{Z}^{(l)}_{fb,i}], \quad \text{where} * \in \{img, gene\}, \quad (6)$$

$$\alpha^{(l)}_{*,ij} = \frac{\exp(w^{(l)}_{att}\sigma(W^{(l)}[h^{(l)}_{*,i} || h^{(l)}_{*,j}]))}{\sum_{k \in \mathcal{N}_i} \exp(w^{(l)}_{att}\sigma(W^{(l)}[h^{(l)}_{*,i} || h^{(l)}_{*,k}]))}, \quad (7)$$

$$\mathcal{Z}^{(l+1)}_{*,i} = \sigma\Big(\sum_{j \in \mathcal{N}_i} \alpha^{(l)}_{*,ij} W^{(l)} h^{(l)}_{*,j}\Big), \quad (8)$$

where σ denotes LeakyReLU, $w^{(l)}_{att} \in \mathbb{R}^D$ and $W^{(l)} \in \mathbb{R}^{D \times (D+D')}$ denote the attention weight matrix and regular weight matrix of the l-th MGDAT block, respectively.

The histology image patches of V_i are reconstructed from the final image embeddings, $\mathcal{Z}_{img,i}$, through a ResNet-based deconvolutional decoder D_2, while the gene expression profiles of V_i are reconstructed from the final gene embeddings, $\mathcal{Z}_{gene,i}$, through a single-layer GNN-based decoder D_3 (Hou et al. 2023):

$$\widetilde{\boldsymbol{P}}_i := D_2(\mathcal{Z}_{img,i}), \quad \widetilde{\boldsymbol{x}}_i := D_3(\mathcal{Z}_{gene,i}) \quad (9)$$

The training loss of stage II includes an image-level loss, same as that defined in Equation (4), and a gene-level reconstruction loss measured in scaled cosine errors:

$$\begin{aligned}\mathcal{L}_{rec} =& \alpha \cdot \sum_i^N (-\text{SSIM}(\boldsymbol{P}_i, \widetilde{\boldsymbol{P}}_i) + ||\boldsymbol{P}_i - \widetilde{\boldsymbol{P}}_i||_1) \\ &+ (1-\alpha) \cdot \sum_i^N \mathcal{L}_{SCE}(\boldsymbol{x}_i, \widetilde{\boldsymbol{x}}_i),\end{aligned} \quad (10)$$

Target Dataset	Metric	Method									
		Multimodal-based		Image-based			ST-based				
		MEATRD	M3DM	SimpleNet	f-AnoGAN	Patch SVDD	DOMINANT	PREM	Spatial-ID	scmap	CAMLU
10x-hBC-A1	AUC	$\mathbf{0.756}_{\pm 0.007}$	$0.520_{\pm 0.046}$	$0.543_{\pm 0.095}$	$\underline{0.642}_{\pm 0.109}$	$0.614_{\pm 0.005}$	$0.488_{\pm 0.117}$	$0.211_{\pm 0.004}$	$0.463_{\pm 0.067}$	$0.500_{\pm 0.000}$	$0.516_{\pm 0.021}$
	F1	$0.892_{\pm 0.007}$	$0.875_{\pm 0.0013}$	$0.875_{\pm 0.011}$	$0.892_{\pm 0.017}$	$\underline{0.892}_{\pm 0.005}$	$0.885_{\pm 0.017}$	$0.865_{\pm 0.000}$	$0.870_{\pm 0.004}$	$\mathbf{0.934}_{\pm 0.000}$	$0.376_{\pm 0.383}$
10x-hBC-B1	AUC	$\mathbf{0.920}_{\pm 0.028}$	$0.505_{\pm 0.029}$	$0.554_{\pm 0.135}$	$\underline{0.736}_{\pm 0.144}$	$0.442_{\pm 0.025}$	$0.698_{\pm 0.077}$	$0.288_{\pm 0.006}$	$0.195_{\pm 0.083}$	$0.504_{\pm 0.000}$	$0.667_{\pm 0.160}$
	F1	$\mathbf{0.841}_{\pm 0.022}$	$0.210_{\pm 0.027}$	$0.302_{\pm 0.127}$	$\underline{0.568}_{\pm 0.176}$	$0.225_{\pm 0.025}$	$0.352_{\pm 0.143}$	$0.073_{\pm 0.008}$	$0.067_{\pm 0.064}$	$0.354_{\pm 0.000}$	$0.428_{\pm 0.365}$
10x-hBC-C1	AUC	$\mathbf{0.715}_{\pm 0.017}$	$0.540_{\pm 0.034}$	$0.501_{\pm 0.099}$	$0.485_{\pm 0.035}$	$0.401_{\pm 0.0032}$	$0.633_{\pm 0.101}$	$0.419_{\pm 0.004}$	$0.384_{\pm 0.055}$	$0.500_{\pm 0.000}$	$\underline{0.660}_{\pm 0.156}$
	F1	$\mathbf{0.842}_{\pm 0.021}$	$0.735_{\pm 0.028}$	$0.735_{\pm 0.024}$	$0.713_{\pm 0.021}$	$0.661_{\pm 0.005}$	$0.769_{\pm 0.040}$	$0.695_{\pm 0.006}$	$0.687_{\pm 0.013}$	$\underline{0.838}_{\pm 0.000}$	$0.808_{\pm 0.021}$
10x-hBC-D1	AUC	$\mathbf{0.803}_{\pm 0.017}$	$0.488_{\pm 0.011}$	$0.485_{\pm 0.111}$	$0.276_{\pm 0.072}$	$0.377_{\pm 0.005}$	$0.530_{\pm 0.172}$	$0.380_{\pm 0.003}$	$0.469_{\pm 0.007}$	$0.503_{\pm 0.000}$	$\underline{0.649}_{\pm 0.066}$
	F1	$\mathbf{0.698}_{\pm 0.016}$	$0.443_{\pm 0.017}$	$0.433_{\pm 0.072}$	$0.253_{\pm 0.085}$	$0.373_{\pm 0.010}$	$0.478_{\pm 0.123}$	$0.344_{\pm 0.010}$	$0.410_{\pm 0.011}$	$\underline{0.626}_{\pm 0.000}$	$0.465_{\pm 0.158}$
10x-hBC-E1	AUC	$\mathbf{0.553}_{\pm 0.046}$	$\underline{0.536}_{\pm 0.014}$	$0.465_{\pm 0.119}$	$0.369_{\pm 0.034}$	$0.300_{\pm 0.009}$	$0.475_{\pm 0.083}$	$0.429_{\pm 0.006}$	$0.449_{\pm 0.082}$	$0.500_{\pm 0.000}$	$0.405_{\pm 0.047}$
	F1	$\mathbf{0.739}_{\pm 0.029}$	$0.598_{\pm 0.009}$	$0.542_{\pm 0.077}$	$0.492_{\pm 0.021}$	$0.443_{\pm 0.006}$	$0.570_{\pm 0.058}$	$0.528_{\pm 0.008}$	$0.542_{\pm 0.054}$	$\underline{0.734}_{\pm 0.000}$	$0.081_{\pm 0.095}$
10x-hBC-F1	AUC	$\mathbf{0.667}_{\pm 0.009}$	$0.485_{\pm 0.046}$	$0.476_{\pm 0.017}$	$0.493_{\pm 0.011}$	$0.483_{\pm 0.005}$	$0.477_{\pm 0.074}$	$0.379_{\pm 0.004}$	$0.380_{\pm 0.074}$	$\underline{0.500}_{\pm 0.000}$	$0.409_{\pm 0.051}$
	F1	$\underline{0.858}_{\pm 0.003}$	$0.832_{\pm 0.009}$	$0.835_{\pm 0.002}$	$0.842_{\pm 0.004}$	$0.840_{\pm 0.003}$	$0.834_{\pm 0.018}$	$0.825_{\pm 0.001}$	$0.820_{\pm 0.005}$	$\mathbf{0.910}_{\pm 0.000}$	$0.036_{\pm 0.022}$
10x-hBC-G2	AUC	$\mathbf{0.640}_{\pm 0.079}$	$0.524_{\pm 0.016}$	$0.482_{\pm 0.074}$	$0.457_{\pm 0.016}$	$0.430_{\pm 0.008}$	$\underline{0.576}_{\pm 0.107}$	$0.430_{\pm 0.006}$	$0.312_{\pm 0.024}$	$0.500_{\pm 0.000}$	$0.518_{\pm 0.001}$
	F1	$\mathbf{0.544}_{\pm 0.045}$	$0.366_{\pm 0.016}$	$0.333_{\pm 0.068}$	$0.295_{\pm 0.002}$	$0.294_{\pm 0.018}$	$0.434_{\pm 0.095}$	$0.273_{\pm 0.006}$	$0.214_{\pm 0.029}$	$\underline{0.510}_{\pm 0.000}$	$0.070_{\pm 0.005}$
10x-hBC-H1	AUC	$\mathbf{0.732}_{\pm 0.064}$	$0.474_{\pm 0.023}$	$0.443_{\pm 0.099}$	$\underline{0.625}_{\pm 0.083}$	$0.415_{\pm 0.005}$	$0.521_{\pm 0.105}$	$0.370_{\pm 0.009}$	$0.319_{\pm 0.061}$	$0.500_{\pm 0.000}$	$0.515_{\pm 0.010}$
	F1	$\mathbf{0.516}_{\pm 0.029}$	$0.273_{\pm 0.029}$	$0.186_{\pm 0.080}$	$0.359_{\pm 0.080}$	$0.066_{\pm 0.003}$	$0.297_{\pm 0.060}$	$0.209_{\pm 0.018}$	$0.179_{\pm 0.038}$	$\underline{0.467}_{\pm 0.000}$	$0.418_{\pm 0.113}$
Mean	AUC	**0.723**	0.509	0.494	0.510	0.433	<u>0.550</u>	0.363	0.371	0.501	0.542
	F1	**0.741**	0.542	0.530	0.552	0.474	0.577	0.476	0.474	<u>0.672</u>	0.335

Table 1: Performance evaluation of anomalous tissue region detection across eight human breast cancer ST datasets. The table presents the results in terms of AUC and F1 scores, with each cell showing the average score from five independent runs and the corresponding standard deviation. The best score for each dataset is **bolded**, and the second-best score is <u>underline</u>.

$$\mathcal{L}_{SCE}(\boldsymbol{x}_i, \widetilde{\boldsymbol{x}}_i) = \left(1 - \frac{\boldsymbol{x}_i^\top \widetilde{\boldsymbol{x}}_i}{||\boldsymbol{x}_i|| \cdot ||\widetilde{\boldsymbol{x}}_i||}\right)^\gamma, \gamma \geq 1, \quad (11)$$

where $0 < \alpha < 1$ is the weight assigned to image reconstruction loss, γ is a scaling factor. The workflow of Stage II is illustrated in Figure 2 and Algorithm 1 in supplementary material C.

Latent Multimodal Reconstruction Loss-based Anomaly Discriminator (Stage III)

Following Stage II, the original and reconstructed image patches of any spot i are processed by a ResNet to generate their respective latent manifolds, denoted as $e_{img,i} \coloneqq \mathrm{ResNet}(\boldsymbol{P}_i)$ and $\tilde{e}_{img,i} \coloneqq \mathrm{ResNet}(\widetilde{\boldsymbol{P}}_i)$, respectively. Here, we employ a light-weight ResNet as the encoder since this stage focuses on calculating latent loss rather than for the more involved tissue image reconstruction task. Similarly, the manifolds of the original and reconstructed gene expression profiles of spot i are generated by an MLP, denoted as $e_{gene,i} \coloneqq \mathrm{MLP}(\boldsymbol{x}_i)$ and $\tilde{e}_{gene,i} \coloneqq \mathrm{MLP}(\widetilde{\boldsymbol{x}}_i)$, respectively. Next, these manifolds are normalized, and a feed-forward network (FFN) maps their weighted averages to a latent space where the multimodal reconstruction error, $\ell_{rec,i}$, is calculated as follows:

$$\boldsymbol{Z}_{fused,i} = \mathrm{FFN}\left(\beta \cdot \frac{e_{img,i}}{||e_{img,i}||} + (1-\beta) \cdot \frac{e_{gene,i}}{||e_{gene,i}||}\right) \quad (12)$$

$$\widetilde{\boldsymbol{Z}}_{fused,i} = \mathrm{FFN}\left(\beta \cdot \frac{\tilde{e}_{img,i}}{||\tilde{e}_{img,i}||} + (1-\beta) \cdot \frac{\tilde{e}_{gene,i}}{||\tilde{e}_{gene,i}||}\right) \quad (13)$$

$$\ell_{rec,i} = \boldsymbol{Z}_{fused,i} - \widetilde{\boldsymbol{Z}}_{fused,i} \quad (14)$$

where $0 < \beta < 1$ represents the relative weight assigned to the histology image. We then train a one-class classifier to collapse latent reconstruction errors of inliers into a compact hypersphere using the loss function:

$$\mathcal{L}_{occ} = \|\ell_{rec,i} - c\|^2 \quad (15)$$

where $c = \frac{1}{N}\sum_{k=1}^{N} \ell_{rec,k}$. The training workflow of Stage III is also illustrated in Algorithm 2 of supplementary material C. At inference time, the anomaly score (AS) of a query spot j is computed as the distance of its latent reconstruction loss to c:

$$AS_j \coloneqq \|\ell_{rec,j} - c\|^2 \quad (16)$$

Given the observation that a gap exists between anomaly scores of inliers and anomalies (Figure 1 in supplementary material B), the AS threshold for discriminating inliers and anomalies is automatically determined using a Maximum A Posteriori-Expectation-Maximization (MAP-EM)-based mixture model, as detailed in supplementary material B.

Experiments

Experimental Settings

Datasets. MEATRD is extensively evaluated across eight breast cancer datasets and four primary sclerosing cholangitis (PSC) datasets. (see supplementary material E for data description and preprocessing).

Baselines. We select nine SOTA image-based, ST-based, and multi-modal AD methods as baselines. Image-based methods include two one-class classification methods, Patch SVDD (Yi and Yoon 2020) and SimpleNet (Liu et al. 2023), alongside a reconstruction-based method, f-AnoGAN (Schlegl et al. 2019). For ST-based methods, we consider scmap (Kiselev, Yiu, and Hemberg 2018), a classification-based method, CAMLU (Li et al. 2022), a reconstruction-based method, PREM (Pan et al. 2023), a discriminative graph method,

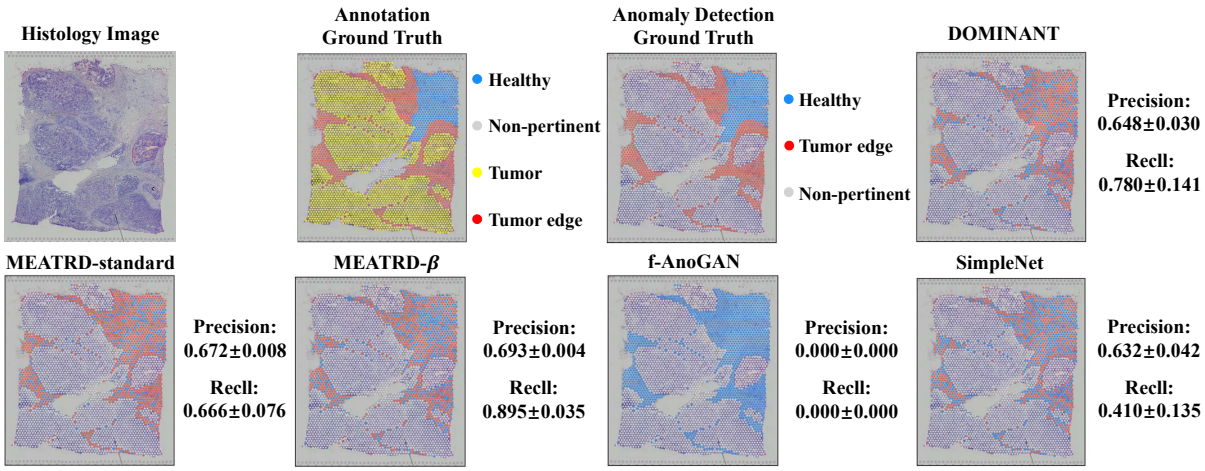

Figure 3: Visualized detection results of tumor edge regions that visually resemble the adjacent normal tissues in the 10x-hBC-I1 dataset. The first row, from left to right, displays the original histology image, the one annotated with ground truth region labels, the one highlighting the tumor edge region (in red) and the adjacent healthy region (in blue), and the one annotated with ATRs identified by DOMINANT. The second row presents images annotated with ATRs identified by their respective methods. The performance of each method is also quantified using mean precision and recall scores over five independent runs. These metrics, along with their standard deviations, are displayed right to each method's panel.

DOMINANT (Ding et al. 2019), a generative graph method, and Spatial-ID (Shen et al. 2022), a classification-based method tailored for ST data. Additionally, M3DM (Wang et al. 2023) is chosen as a representative multimodal baseline.

Evaluation Protocols. AUC and F1 scores are used to evaluate the accuracy of ATR detection. For a fair comparison, the F1 score is calculated with the threshold matching the actual proportion of true anomalies (Shenkar and Wolf 2021). Reported metrics and standard deviations are averaged over five independent runs.

Anomalous Tissue Region Detection

In this experiment, as listed in supplementary material F, MEATRD is trained on eight human normal breast ST datasets (i.e., 10x-hNB-{v03-v10}) and tested on eight human breast cancer (i.e., 10x-hBC-{A1-H1}) ST datasets. Table 1 showcases MEATRD's superiority over baselines in detecting ATRs across datasets, consistently ranking first in AUC scores and six times in F1 scores. It outperforms the second-best performing method with an average leap of 17.45% in AUC scores and 10.31% in F1 scores. Furthermore, Table 4 in supplementary material F indicates that our model performed well in detecting PSCs, demonstrating its generalization to other types of diseases. Generally, methods that use ST data, for example, DOMINANT, scmap and CAMLU, tend to outperform those that rely solely on histology images, indicating the pivotal role of gene expression information provided by ST data in aiding the detection of ATRs, especially those visually similar to normal tissues. Moreover, we find that, DOMINANT, a graph-based AD method, prevails over other baselines, and that M3DM, a multimodal method that utilizes both image and ST data yet fails to account for spatial relationships among spots, does not perform as well as MEATRD. These observations emphasize the value of spatial contextual information in accurate ATR detection.

Discovering Anomalous Tissue Regions Visually Similar to Normal Tissues

To evaluate the efficacy of MEATRD in detecting ATRs with minimal visual distinctions from normal tissues, we conduct a comparative analysis on the 10x-hBC-I1 ST dataset, which encompasses a tumor edge region that visually blends with the adjacent normal tissues, as indicated in red in the annotated histology image from Figure 3. Our analysis includes: the standard MEATRD implementation (MEATRD-standard); MEATRD-β, a variant that downplays the influence of histology image by decreasing β from 0.5 to 0.1 in Equation (12) and Equation (13); DOMINANT, the top performing baseline utilizing ST data from the previous section; two leading image-based AD methods, f-AnoGAN and SimpleNet. The results, visually presented in Figure 3 demonstrate that MEATRD-β more accurately identifies spots within the tumor edge region as anomalous, compared to the other competing methods. This finding is quantitatively supported by its highest precision (0.693) and recall (0.895) scores. The observation that MEATRD-standard, MEATRD-β, and DOMINANT prevail over the image-based AD methods underscores the value of using ST data for pinpointing pathogenic tissue regions that visually resemble normal tissues. Furthermore, DOMINANT's marginal performance edge over MEATRD-standard suggests that in this specific context, the histology image contributes very limited additional information. Indeed, MEATRD-β, which places greater emphasis on ST data, showcases an improved performance of 3.1% in precision and 34.4% in recall, compared to MEATRD-standard. Nonetheless, for scenarios involving low-quality ST data and visually traceable ATRs, incorpo-

Metric	Parameter α			Parameter β			Embedding dimension			Bottleneck dimension		
	0.9	0.5	0.1	0.9	0.5	0.1	128	256	512	16	64	256
AUC	0.678	**0.723**	0.709	0.654	**0.723**	0.699	0.705	**0.723**	0.721	**0.723**	0.715	0.682
F1	0.696	**0.741**	0.725	0.668	**0.741**	0.718	0.726	**0.741**	0.735	**0.741**	0.728	0.711
Metric	Detection dimension			MGDAT Layers			MGDAT Layers					
	64	128	256	2	3	4	1	2	4			
AUC	0.606	0.720	**0.723**	0.694	**0.723**	0.533	0.718	**0.723**	0.721			
F1	0.623	0.732	**0.741**	0.719	**0.741**	0.565	0.730	**0.741**	0.737			

Table 2: Sensitivity analysis of hyperparameter in MEATRD across eight human breast cancer datasets. Default settings are marked in gray.

Metric	Ablation study						
	ST only	Image only	w/o MGDAT	w/o TNM	w/o RE	w/o OC	Full
AUC	0.631	0.497	0.639	0.655	0.642	0.584	**0.723**
F1	0.667	0.544	0.689	0.699	0.685	0.631	**0.741**

Table 3: Ablation study of key components in MEATRD across eight human breast cancer datasets. Method performance is gauged through average AUC and F1 scores. "Full" represents the complete MEATRD model. "ST Only" and "Image Only" utilize only ST data or histology images, respectively. "w/o MGDAT" omits the MGDAT block. "w/o TNM" omits the target-node-masking technique. "w/o RE" substitutes the latent multimodal reconstruction errors with direct spot embeddings for input to the discriminative model in Stage III. "w/o OC" discards the entire stage III and utilizes spot reconstruction errors as anomaly scores for ATR detection.

rating visual cues from histology images are undoubtedly beneficial, as established in our prior analysis and ablation study.

Ablation Studies

We conduct ablation studies over the eight human breast cancer ST datasets (i.e., 10x-hBC-{A1-H1}) to investigate the effects of MEATRD's key components in ATR detection. These components include using multiple data modality, multimodal data fusion using fused bottleneck embedding, masking for target node reconstruction, multimodal reconstruction losses in the one-class classifier in Stage III, enlarging anomaly score discrepancy between inliers and anomalies using a one-class classifier, using Mobile-Unet as the pretraining backbone in Stage I. The descriptions detailed in the *Ablation Studies* section in supplementary material F, demonstrate that removing any of these components leads to suboptimal performance. This is due to the inefficient use of cross-modal complementary information, less effective addressing of model over-generalization, and increased sensitivity to reference-target domain shifts.

Sensitivity Analysis

Here, we conduct sensitivity analyses on eight 10x-hBC datasets to examine the effects of MEATRD's key hyperparameters, including α and β, which control the relative weights between gene and image modalities in Stage II and III; the dimensions of visual and gene embedding from Stage I, bottleneck embedding in Stage II, and the inputs to the one-classification classifier in Stage III; the number of MGDAT layers and its attention heads. The effect of these parameters on MEATRD's performance, measured by AUC and F1 scores, is presented in Table 2. Detailed analyses are provided in supplementary material F.3.

Complexity Analysis

We analyze the model complexity of MEATRD across its three stages by evaluating the number of parameters, computational performance (MFlops), time complexity, training time, and inference time. We also compare these metrics with the nine baseline methods. Detailed results are provided in supplementary material F.4. In summary, MEATRD is scalable to the number of spots and edges (proportional to the number of spots due to the adjacency matrix setting) and demonstrates good efficiency in our experiments.

Conclusion

In this paper, we propose MEATRD, a pilot method that integrates histology images and ST data to enhance ATR detection at both visual and molecular levels. MEATRD treats tissue spots as nodes in an attributed graph to embed spatial relationships into their representations. The MGDAT network, a key innovation of MEATRD, facilitates effective cross-node and cross-modality information exchange, enabling comprehensive graph representation learning. MEATRD harmonizes one-class classification with reconstruction deviation-based AD detection, simultaneously addressing the challenges of reference-target domain shift and model over-generalization. Rigorous evaluations on a suite of real ST datasets have demonstrated MEATRD's superiority over various SOTA AD methods in detecting ATRs including those that are visually akin to contextual normal tissues. Furthermore, MEATRD also offers a framework generalizable to other multimodal AD tasks involving compatible imagery and graph data modalities.

Acknowledgments

The project is funded by the Excellent Young Scientist Fund of Wuhan City (Grant No. 21129040740) to X.S.

References

Bejnordi, B. E.; Veta, M.; Van Diest, P. J.; Van Ginneken, B.; Karssemeijer, N.; Litjens, G.; Van Der Laak, J. A.; Hermsen, M.; Manson, Q. F.; Balkenhol, M.; et al. 2017. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22): 2199–2210.

Bergmann, P.; Fauser, M.; Sattlegger, D.; and Steger, C. 2019. MVTec AD – A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Deng, H.; and Li, X. 2022. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9737–9746.

Ding, K.; Li, J.; Bhanushali, R.; and Liu, H. 2019. Deep anomaly detection on attributed networks. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, 594–602. SIAM.

Donahue, J.; Krähenbühl, P.; and Darrell, T. 2016. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*.

Dong, K.; and Zhang, S. 2022. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nature communications*, 13(1): 1739.

Hou, Z.; He, Y.; Cen, Y.; Liu, X.; Dong, Y.; Kharlamov, E.; and Tang, J. 2023. GraphMAE2: A Decoding-Enhanced Masked Self-Supervised Graph Learner. In *Proceedings of the ACM Web Conference 2023*, 737–746.

Hu, J.; Coleman, K.; Zhang, D.; Lee, E. B.; Kadara, H.; Wang, L.; and Li, M. 2023. Deciphering tumor ecosystems at super resolution from spatial transcriptomics with TESLA. *Cell systems*, 14(5): 404–417.

Hu, J.; Li, X.; Coleman, K.; Schroeder, A.; Ma, N.; Irwin, D. J.; Lee, E. B.; Shinohara, R. T.; and Li, M. 2021. SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nature methods*, 18(11): 1342–1351.

Kiselev, V. Y.; Yiu, A.; and Hemberg, M. 2018. scmap: projection of single-cell RNA-seq data across data sets. *Nature methods*, 15(5): 359–362.

Komura, D.; and Ishikawa, S. 2018. Machine learning methods for histopathological image analysis. *Computational and structural biotechnology journal*, 16: 34–42.

Li, Z.; Wang, Y.; Ganan-Gomez, I.; Colla, S.; and Do, K.-A. 2022. A machine learning-based method for automatically identifying novel cells in annotating single-cell RNA-seq data. *Bioinformatics*, 38(21): 4885–4892.

Liu, Z.; Zhou, Y.; Xu, Y.; and Wang, Z. 2023. Simplenet: A simple network for image anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20402–20411.

Moses, L.; and Pachter, L. 2022. Museum of spatial transcriptomics. *Nat Methods*, 19: 534–546.

Nagrani, A.; Yang, S.; Arnab, A.; Jansen, A.; Schmid, C.; and Sun, C. 2021. Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems*, 34: 14200–14213.

Okada, M.; and Taniguchi, T. 2021. Dreaming: Model-based reinforcement learning by latent imagination without reconstruction. In *2021 ieee international conference on robotics and automation (icra)*, 4209–4215. IEEE.

Ouardini, K.; Yang, H.; Unnikrishnan, B.; Romain, M.; Garcin, C.; Zenati, H.; Campbell, J. P.; Chiang, M. F.; Kalpathy-Cramer, J.; Chandrasekhar, V.; et al. 2019. Towards practical unsupervised anomaly detection on retinal images. In *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data: First MICCAI Workshop, DART 2019, and First International Workshop, MIL3ID 2019, Shenzhen, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13 and 17, 2019, Proceedings 1*, 225–234. Springer.

Pan, J.; Liu, Y.; Zheng, Y.; and Pan, S. 2023. PREM: A Simple Yet Effective Approach for Node-Level Graph Anomaly Detection. *arXiv preprint arXiv:2310.11676*.

Pocevičiūtė, M.; Eilertsen, G.; and Lundström, C. 2021. Unsupervised anomaly detection in digital pathology using GANs. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 1878–1882. IEEE.

Riasatian, A.; Babaie, M.; Maleki, D.; Kalra, S.; Valipour, M.; Hemati, S.; Zaveri, M.; Safarpoor, A.; Shafiei, S.; Afshari, M.; et al. 2021. Fine-tuning and training of densenet for histopathology image representation using tcga diagnostic slides. *Medical Image Analysis*, 70: 102032.

Ristea, N.-C.; Madan, N.; Ionescu, R. T.; Nasrollahi, K.; Khan, F. S.; Moeslund, T. B.; and Shah, M. 2022. Self-supervised predictive convolutional attentive block for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13576–13586.

Roth, K.; Pemula, L.; Zepeda, J.; Schölkopf, B.; Brox, T.; and Gehler, P. 2022. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14318–14328.

Ruff, L.; Vandermeulen, R.; Goernitz, N.; Deecke, L.; Siddiqui, S. A.; Binder, A.; Müller, E.; and Kloft, M. 2018. Deep one-class classification. In *International conference on Machine Learning*, 4393–4402. PMLR.

Schlegl, T.; Seeböck, P.; Waldstein, S. M.; Langs, G.; and Schmidt-Erfurth, U. 2019. f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. *Medical image analysis*, 54: 30–44.

Shen, R.; Liu, L.; Wu, Z.; Zhang, Y.; Yuan, Z.; Guo, J.; Yang, F.; Zhang, C.; Chen, B.; Feng, W.; et al. 2022. Spatial-ID: a cell typing method for spatially resolved transcriptomics via transfer learning and spatial embedding. *Nature communications*, 13(1): 7640.

Shenkar, T.; and Wolf, L. 2021. Anomaly detection for tabular data with internal contrastive learning. In *International Conference on Learning Representations*.

Shenkar, T.; and Wolf, L. 2022. Anomaly detection for tabular data with internal contrastive learning. In *International conference on learning representations*.

Shvetsova, N.; Bakker, B.; Fedulova, I.; Schulz, H.; and Dylov, D. V. 2021. Anomaly detection in medical imaging with deep perceptual autoencoders. *IEEE Access*, 9: 118571–118583.

Sohn, K.; Li, C.-L.; Yoon, J.; Jin, M.; and Pfister, T. 2020. Learning and evaluating representations for deep one-class classification. *arXiv preprint arXiv:2011.02578*.

Srinidhi, C. L.; Ciga, O.; and Martel, A. L. 2021. Deep neural network models for computational histopathology: A survey. *Medical image analysis*, 67: 101813.

Tschuchnig, M. E.; and Gadermayr, M. 2022. Anomaly detection in medical imaging-a mini review. In *Data Science–Analytics and Applications: Proceedings of the 4th International Data Science Conference–iDSC2021*, 33–38. Springer.

Wang, G.; Wu, S.; Xiong, Z.; Qu, H.; Fang, X.; and Bao, Y. 2024. CROST: a comprehensive repository of spatial transcriptomics. *Nucleic Acids Research*, 52(D1): D882–D890.

Wang, L.; Maletic-Savatic, M.; and Liu, Z. 2022. Region-specific denoising identifies spatial co-expression patterns and intra-tissue heterogeneity in spatially resolved transcriptomics data. *Nature Communications*, 13(1): 6912.

Wang, Y.; Peng, J.; Zhang, J.; Yi, R.; Wang, Y.; and Wang, C. 2023. Multimodal Industrial Anomaly Detection via Hybrid Fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8032–8041.

Wang, Y.; Song, B.; Wang, S.; Chen, M.; Xie, Y.; Xiao, G.; Wang, L.; and Wang, T. 2022. Sprod for de-noising spatially resolved transcriptomics data based on position and image information. *Nature methods*, 19(8): 950–958.

Xu, K.; Lu, Y.; Hou, S.; Liu, K.; Du, Y.; Huang, M.; Feng, H.; Wu, H.; and Sun, X. 2024. Detecting anomalous anatomic regions in spatial transcriptomics with STANDS. *Nature Communications*, 15(1): 8223.

Yi, J.; and Yoon, S. 2020. Patch svdd: Patch-level svdd for anomaly detection and segmentation. In *Proceedings of the Asian conference on computer vision*.

Zehnder, P.; Feng, J.; Fuji, R. N.; Sullivan, R.; and Hu, F. 2022. Multiscale generative model using regularized skip-connections and perceptual loss for anomaly detection in toxicologic histopathology. *Journal of Pathology Informatics*, 13: 100102.

Zingman, I.; Stierstorfer, B.; Lempp, C.; and Heinemann, F. 2023. Learning image representations for anomaly detection: application to discovery of histological alterations in drug development. *Medical Image Analysis*, 103067.

Zong, Y.; Yu, T.; Wang, X.; Wang, Y.; Hu, Z.; and Li, Y. 2022. conST: an interpretable multi-modal contrastive learning framework for spatial transcriptomics. *bioRxiv*, 2022–01.