

Gene-DML: Dual-Pathway Multi-Level Discrimination for Gene Expression Prediction from Histopathology Images

Yaxuan Song¹, Jianan Fan^{1*}, Hang Chang², Weidong Cai^{1*}

¹The University of Sydney, Australia ²Lawrence Berkeley National Laboratory, USA

yson2999@sydney.edu.au, jianan.fan@sydney.edu.au, hchang@lbl.gov, tom.cai@sydney.edu.au

Abstract

*Accurately predicting gene expression from histopathology images offers a scalable and non-invasive approach to molecular profiling, with significant implications for precision medicine and computational pathology. However, existing methods often underutilize the cross-modal representation alignment between histopathology images and gene expression profiles across multiple representational levels, thereby limiting their prediction performance. To address this, we propose **Gene-DML**, a unified framework that structures latent space through **Dual-pathway Multi-Level discrimination** to enhance correspondence between morphological and transcriptional modalities. The **multi-scale instance-level** discrimination pathway aligns hierarchical histopathology representations extracted at local, neighbor, and global levels with gene expression profiles, capturing scale-aware morphological-transcriptional relationships. In parallel, the **cross-level instance-group** discrimination pathway enforces structural consistency between individual (image/gene) instances and modality-crossed (gene/image, respectively) groups, strengthening the alignment across modalities. By jointly modeling fine-grained and structural-level discrimination, **Gene-DML** is able to learn robust cross-modal representations, enhancing both predictive accuracy and generalization across diverse biological contexts. Extensive experiments on public spatial transcriptomics datasets demonstrate that **Gene-DML** achieves state-of-the-art performance in gene expression prediction. The code and processed datasets are available at <https://github.com/YXSong000/Gene-DML>.*

1. Introduction

Spatial transcriptomics (ST) has emerged as a transformative technology that enables high-resolution mapping of quantification of mRNA expression (gene expression) within intact tissue architectures [27, 29, 38]. Unlike

bulk RNA sequencing [35], single-cell RNA sequencing (scRNA-seq) or single-nucleus RNA sequencing (snRNA-seq), which lacks spatial localization, spatial transcriptomics preserves the anatomical and microenvironmental context of gene expressions, making it invaluable for providing unprecedented insights into the spatial organization of molecular processes [41]. Intuitively, ST technology can be seen as converging on a gene expression matrix that captures the transcriptome at every spot (that is, a pixel, a cell, or a group of cells) [27], which is divided from large-scale whole-slide images (WSIs). Traditional ST methods, such as Visium [29], MERFISH [4], seqFISH+ [9], STARmap [36] *et al.* [13], tend to be time-consuming and expensive, relying on specialized instruments and extensive domain knowledge. This raises a critical question: how can we leverage recent advances in deep learning to accurately infer spatial gene expression directly from WSIs?

Several recent studies have made substantial progress in gene expression prediction from WSIs [23, 43]. Unimodal approaches [7, 14, 25, 42, 45] typically employ supervised learning to predict gene profiles from WSI tiles, focusing on improving feature extraction to learn strong visual representations. However, despite the use of powerful image encoders, these unimodal methods often fail to incorporate gene-specific morphological information. The absence of alignment with gene expression weakens the model's ability to capture meaningful biological signals, often resulting in overfitting to high-dimensional gene expression counts and weak generalization across patient samples.

To explore image-gene relationships more effectively, bimodal methods [17, 22, 24] co-embed WSI tiles and gene expression profiles in a shared space. While these approaches enable cross-modal learning, they typically model only single-level, pairwise relationships between individual tiles and gene instances. By treating each image-gene pair independently, they fail to capture the hierarchical alignment inherent in neither WSI tile nor gene expression profile instances, triggering overfitting to the instance-level noise. Thus, multi-level alignments between image and gene modalities are essential for constructing a ro-

*Corresponding authors.

bust and expressive latent space that faithfully encodes both histopathology morphology [11, 28, 40] and gene expression; otherwise, the model’s generalization and the learning of biologically coherent representations are compromised.

To this end, we propose **Gene-DML**, a novel framework that enhances **Gene**-image correspondence by performing **Dual-pathway Multi-Level** discrimination. It jointly captures the scale-aware morphological-transcriptional relationship between multi-scaled WSI tiles and gene expression profiles at the instance level and instance-group level, that is, WSI tile instances aligned with clustered gene-expression groups and vice versa. By jointly modeling fine-grained and high-level discrimination, Gene-DML enables the learning of robust cross-modal representations, enhancing both predictive accuracy and generalization across diverse biological contexts. Contributions are summarized as:

- We propose Gene-DML, a novel framework that models dual-pathway multi-level discrimination to enhance spatial and morphological alignment between histopathology images and gene expression profiles.
- Gene-DML integrates multi-scale instance-level discrimination pathway to align gene expression with WSI features extracted at local, neighbor, and global scales, enabling scale-aware cross-modal representation learning.
- In parallel, Gene-DML incorporates the cross-level instance-group discrimination pathway, which jointly enforces consistency of instance-level and group-level across image and gene modalities, improving semantic structure in the shared latent space.
- Through comprehensive experiments on public ST datasets and additional external tests on Visium data, Gene-DML outperforms state-of-the-art methods on multiple datasets, achieving superior prediction accuracy and generalization across heterogeneous patient samples.

2. Related Work

2.1. Gene Expression Prediction

Early methods for spatial gene expression prediction from histology [10], such as ST-Net [14], employed a transfer learning strategy by fine-tuning a DenseNet121 model [16] pre-trained on ImageNet to capture image features and map WSI tiles to gene expression labels. Building upon this, HisToGene [25] integrates Vision Transformers (ViT) to capture global contextual dependencies between image tiles within a WSI. Further advancing this direction, Hist2ST [45] incorporates ConvMixer [31] and a graph convolution network (GCN) post-initially for enhanced image tile embeddings by aggregating neighborhood information. Recent multi-scale models, TRIPLEX [7] and M2OST [34] improved performance by integrating spot-, neighborhood-, and slide-level features. These architectures leveraged hierarchical modeling and fusion techniques

to capture both fine-grained morphology and global tissue context. EGN [42] is an alternative paradigm in which image embedding is optimized by selecting the most analogous exemplars from the target spot within a WSI to enhance gene expression prediction. Despite these advances, these unimodal methods lack semantic alignment with gene expression and struggle to generalize across samples. Bimodal contrastive learning has gained traction. To explore the image-gene relationship for more scalable and broader potential approaches, bimodal models, BLEEP [41] and MclSTExp [24] with Transformer encoders align image and gene embeddings through contrastive loss, enabling gene querying via latent similarity. RankByGene [17] and ST-Align [22] further improved alignment through ranking consistency and large-scale pretraining, while Stem [46] explored generative modeling using diffusion techniques [32]. Existing contrastive or generative methods focus mainly on pairwise alignment and overlook deeper morphological-transcriptional correspondences across image scales, which is essential for biological semantic understanding.

2.2. Multi-scaled Image Feature Extraction

Several recent works have explored multi-scale modeling for WSIs to capture both fine-grained and contextual features [12]. HIPT by Chen *et al.* [5] employs hierarchical self-supervised learning to learn patch-level and region-level features via contrastive objectives, enabling effective representation learning across image resolutions. TRIPLEX [7] is proposed by integrating multi-resolution features from target, neighbor, and global tissue levels using a supervised fusion strategy. M2OST [34] leverages multi-scale WSI patches with Transformer-based many-to-one regression for accurate gene prediction. Cross-scale MIL architectures have also shown promise. MEGT [8] and DSF-WSI [33] adopt dual-branch architectures with separate high- and low-resolution encoders, fusing outputs using graph-Transformer modules or masked-jigsaw tasks. Despite these advances, existing methods suffer from the following limitations: (1) most focus exclusively on unimodal representation learning without leveraging alignment of transcriptomic modality data, and (2) fusion-based designs often lack contrastive objectives that explicitly enforce scale-consistent representation. In contrast, Gene-DML incorporates multi-scale instance-level discrimination that explicitly aligns local, neighbor, and global image features with gene expression, enriching cross-modal representations and enhancing semantic consistency across scales.

2.3. Cross-Level Instance-Group Discrimination

Cross-Level Discrimination (CLD) [37] is proposed to improve instance-level contrastive learning by incorporating dynamic group-level structure within a single modality. By pairing each instance with its group centroid, CLD miti-

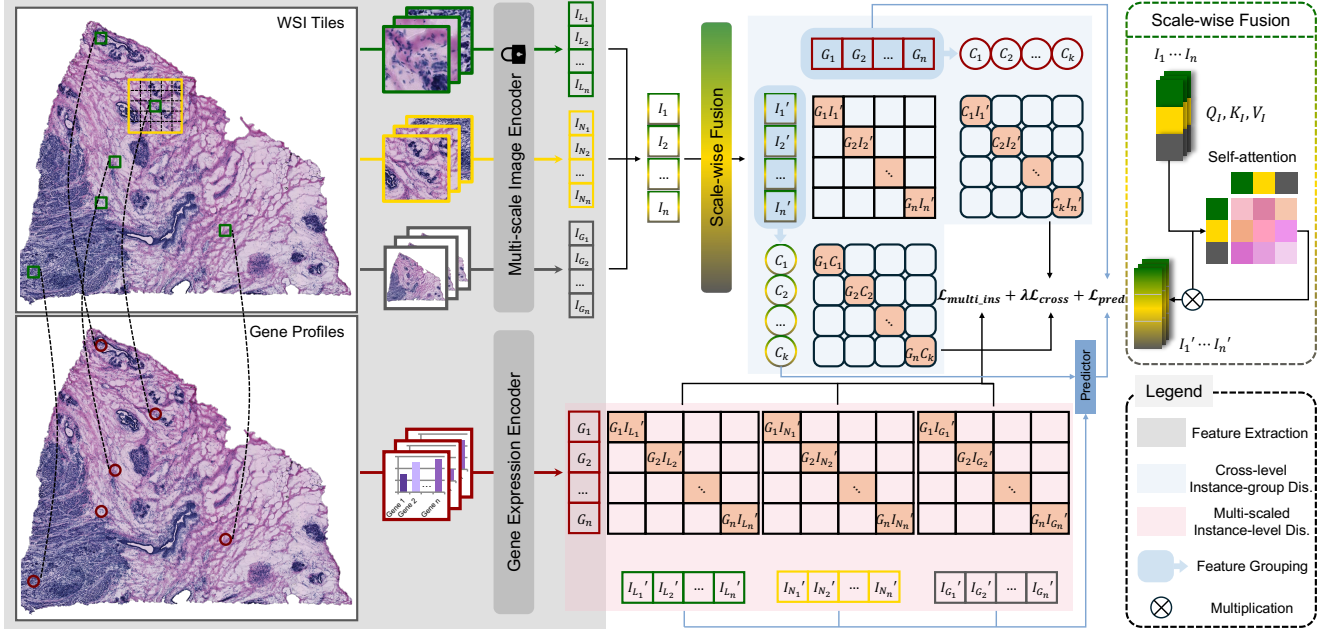


Figure 1. An overview of Gene-DML framework. Pairs of WSI tiles and gene profiles are aligned from both the **multi-scale instance-level** and the **cross-level instance-group** discrimination. An illustration of feature grouping with instance-group alignment is shown in Fig. 2.

gates over-separation of semantically similar samples and improves representation stability. Subsequent work has adapted instance-group techniques within a single modality and for video-text alignment, *e.g.*, CrossCLR [47], TCL [21], by incorporating grouping or de-biasing false negatives, these methods empirically show improved retrieval, classification, and stability [21]. However, they remain unimodal or restricted to intra-modality alignment. In medical imaging and genomics, multimodal frameworks like ContIG [30] and MGI [21] have adopted instance-level contrastive learning across modalities (*e.g.*, image-genetic or image-omics pairs), but these lack mechanisms to model group-level relationships or latent clustering in cross-modal [3]. They typically use pairwise contrastive objectives akin to CLIP [26, 39], which may overemphasize instance-level uniqueness and discount shared biological patterns. To bridge these gaps, Gene-DML integrates a cross-level instance-group discrimination mechanism within two modalities: images and transcriptomics. It aligns each image instance with gene-expression groups and vice versa, ensuring the model learns shared semantic structure between morphology and transcriptional activity. This bidirectional group alignment improves robustness to noise, enhances positive/negative contrastive signal, and supports generalization to unseen tissue domains.

3. Methodology

In this study, we explore how to effectively leverage the alignment information between WSIs and gene expression

profiles via a dual-pathway discrimination strategy to enhance gene expression prediction. Specifically, our objective is to train a bimodal contrastive learning model that captures semantic alignment information between morphological features and gene functions, which is distilled using WSI tiles, represented as $\mathbf{I} \in \mathbb{R}^{N \times H \times W \times 3}$, to predict the gene expression levels for each spot, represented as $\mathbf{G} \in \mathbb{R}^{N \times M}$, in a N -spot WSI. Here, M denotes the number of gene types whose expression levels are to be predicted, and H and W signify the height and width of each spot image, respectively. An overview of the framework we proposed, Gene-DML, is illustrated in Fig. 1.

3.1. Multi-scale Instance-level Alignment

3.1.1. Multi-scale Image Instance Feature

A single spot's gene expression is influenced not only by its WSI tile structure but also by its surrounding environment [7]. Each WSI is split into tiles centered on spatial transcriptomics spots, extracted at three spatial scales: local tiles I_L focus on local-level detail, neighbor tiles I_N capture microenvironment context, and the global scale I_G represents the WSI's global tissue structure. These tiles are encoded via a shared Multi-Scale Image Encoder. Specifically, we utilize the UNI [6] model, pretrained using more than 100 million H&E-stained WSIs, for histopathology image feature extraction. In order to avoid the overfitting problem, the base feature extraction model UNI is kept frozen to ensure the generalization of Gene-DML. The features from all three scales are individually projected into 512 dimensions via separate linear layers.

Local-scale. For local-scale (green square in Fig. 1) feature extraction, the pretrained UNI model is utilized to process N tiles with a size of 224×224 for each WSI. These features are extracted from the forward feature of the UNI model. These features are embedded into 265 features with 1024 dimensions for each.

Neighbor-scale. As for the neighbor-scale (yellow square in Fig. 1), to preserve the structural integrity of the tissue within the WSI, we utilize a region of size 1120×1120 pixel tiles of centered spot rather than the surrounding 5 tiles. These neighbor-scale features are embedded into 25 1024×1024 features by using the UNI model. Since the UNI model is frozen, a series of self-attention blocks are applied on neighbor-scale features, which are kept updated for better model learning.

Global-scale. To efficiently capture global spatial context in WSIs without incurring high computational costs, inspired by [7], the global-scale representation is concatenated with all local-scale tile features extracted by the pretrained UNI model, which is then encoded through the transformer block for the ultimate global-scale token.

Scale-wise Fusion. To effectively integrate morphological information across scales, multi-scale image features captured at local, neighbor, and global scales are concatenated for each spot along the channel dimension $I_n = \{I_{L_n}, I_{N_n}, I_{G_n}\}$ and passed into a Scale-wise Fusion module designed to model cross-scale dependencies. This module adopts a self-attention mechanism that dynamically reweights information across levels, enabling the network to integrate fine-grained cellular patterns with broader tissue context. The fused outputs $I_i^{E_{\text{ins}}} = \{I'_1, I'_2, \dots, I'_n\}$, where $i = 1, \dots, n$, retain scale-specific attributes and are subsequently disentangled for downstream alignment with gene expression features. This process ensures semantically enriched, scale-aware representations for each spot, facilitating more robust cross-modal learning.

3.1.2. Gene Expression Instance Feature

In parallel, to model transcriptomic information in the shared latent space, we employ a lightweight yet expressive fully connected network as the gene expression encoder. Each spatial gene expression profile is passed through the FCN to produce a dense embedding:

$$G_i^{E_{\text{ins}}} = \text{FFN} \circ \text{GeneEncoder}(G_i) \in \mathbb{R}^d, \quad (1)$$

where $i = 1, \dots, n$. Here, GeneEncoder first maps raw gene profiles into an intermediate embedding space of dimension d , capturing high-level transcriptomic structure, consisting of two linear layers with GELU [15] activation and dropout for regularization. This is followed by a FeedForward (FFN) network consisting of two linear layers with a

non-linear activation and dropout, which refines the representation and projects it into the shared latent space. This design allows Gene-DML to learn instance-level gene features that align with corresponding histology-derived representations.

3.2. Cross-Level Instance-Group Alignment

3.2.1. Feature Grouping

To enrich broader semantic alignment information, which enforces the transcriptional activity of a certain tile of WSI close to a group of tiles with similar morphology and vice versa, between the morphological features of WSI tiles and the functional meaning of gene expression, we integrate a **feature grouping** branch for both histopathology and gene expression modalities. As illustrated in the light blue section of Fig. 1, for both feature grouping branches of image and gene, group features are derived from the instance-level embeddings $\{I_i^{E_{\text{ins}}}\}$ and $\{G_i^{E_{\text{ins}}}\}$ (Sec. 3.1) via a fully connected transformation:

$$I_i^{E_{\text{clu}}} = \text{FC}(I_i^{E_{\text{ins}}}) \text{ and } G_i^{E_{\text{clu}}} = \text{FC}(G_i^{E_{\text{ins}}}). \quad (2)$$

This branch serves as a preparatory step for downstream clustering by projecting instance features into a space optimized for group structure discovery. Compared to the instance embeddings ($I_i^{E_{\text{ins}}}$ and $G_i^{E_{\text{ins}}}$) that may contain modality-specific noise or fine-grained variations, the learned group features emphasize higher-level semantic consistency, making them more suitable for clustering. This facilitates more coherent dual-view contrastive alignment not only at the instance level but also across clusters that capture shared morphological or transcriptional patterns.

To structurally organize the latent space and uncover shared biological semantics, we perform modality-specific feature grouping based on the group features $\{I_i^{E_{\text{clu}}}\}$ and $\{G_i^{E_{\text{clu}}}\}$. Specifically, the K -means strategy [44] is applied to unit-normalized vectors to form k local clusters in each modality. This yields sets of centroids $\{C_1^I, \dots, C_k^I\}$ and $\{C_1^G, \dots, C_k^G\}$, which are $\{C_i^I\}$ in gradient-colored circles and $\{C_i^G\}$ in red circles, respectively, as shown in Fig. 1.

3.2.2. Cross-Level Alignment

As for cross-level instance-group alignment, each instance-level embedding $I_i^{E_{\text{ins}}}$ (or $G_i^{E_{\text{ins}}}$) is assigned to its nearest gene (or image) cluster centroid via a mapping $\mathcal{A}(i) = j$. As illustrated in Fig. 2, we then compute instance-to-group similarities bidirectionally across modalities:

- Aligning each gene instance $G_i^{E_{\text{ins}}}$ with its corresponding visual group centroid C_j^I ;
- Aligning each image instance $I_i^{E_{\text{ins}}}$ with its corresponding gene group centroid C_j^G .

This instance-group alignment enables Gene-DML to reinforce semantic consistency across heterogeneous domains

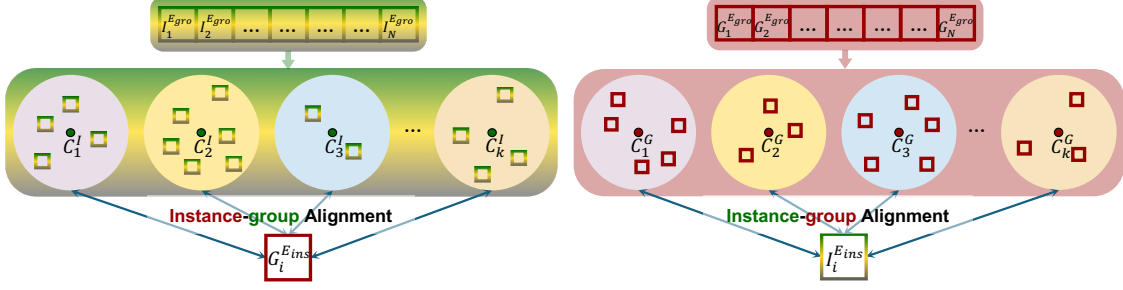


Figure 2. Detailed illustration of feature grouping intuition. The image group features (gradient color squares) and gene group features (red squares) are grouped into k groups with centroids $\{C_1^I, \dots, C_k^I\}$ and $\{C_1^G, \dots, C_k^G\}$.

while mitigating overfitting to noisy or isolated instance-level correlations. By leveraging local group structure in both vision and gene spaces, the model gains robustness and improved generalization to out-of-distribution samples.

3.3. Loss Component

Our method follows the CLIP [26] framework that maximizes the cosine similarity of the image and gene embeddings of the N real pairs while minimizing the cosine similarity of the embeddings of the $N^2 - N$ incorrect pairings at the image-gene instance level. At the multi-scale instance-level, the contrastive loss is applied on WSI embeddings from different scales with gene embedding. At the instance-group level, Gene-DML applies a batch-wise contrastive loss across views between group features I_i^{Eins} and group centroids C_i^G , and vice versa for G_i^{Eins} (with C_i^I), to supervise the consistency of grouping relationships.

3.3.1. Multi-scale Instance-level Loss

Inspired by the contrastive loss used in [41], Gene-DML computes the internal multi-scaled similarities between the image or the gene expression features in advance, denoted as \mathbf{T} , increasing the coherence of the latent space for joint embedding. Then, cross-entropy loss is applied to align the multi-scaled image and gene expression instance features:

$$\mathbf{T}^s = \sigma \left(\frac{\text{sim}(I_s^{Eins}, I_s^{Eins}) + \text{sim}(G_s^{Eins}, G_s^{Eins})}{2 \cdot \tau} \right), \quad (3)$$

$$\mathcal{L}_{\text{multi.ins}} = \frac{1}{3} \sum_s \frac{1}{N} \sum_{i=1}^N \left[- \sum_{j=1}^N \mathbf{T}^s_{ij} \log \sigma \left(\text{sim}(I_{s_i}^{Eins}, G_j^{Eins}) \right) - \sum_{j=1}^N \mathbf{T}^s_{ji} \log \sigma \left(\text{sim}(I_{s_j}^{Eins}, G_i^{Eins}) \right) \right], \quad (4)$$

where s denotes the scale of WSI tiles $s \in \{L, N, G\}$, $\sigma(\cdot)$ denotes the softmax function, and $\text{sim}(A, B) = AB^T$.

3.3.2. Instance-group Loss

The discrimination between instance and group can be regarded as minimizing the cross-entropy between the hard

grouping centroids and the predicted grouping assignment. Specifically, the cross-entropy between the group centroids C_i^G computed by G_i^{Eclu} and the soft grouping assignment predicted from I_i^{Eins} should be minimized, and vice versa for G_i^{Eins} . Thus, we have the instance-group cross-level loss that validates feature grouping across different views:

$$\mathcal{L}_{\text{cross}} = \frac{1}{N} \sum_{i=1}^N \left[- \sum_{j=1}^N \mathbf{T}^I_{ij} \log \sigma \left(\text{sim}(I_i^{Eins}, C_j^G) \right) - \sum_{j=1}^N \mathbf{T}^G_{ji} \log \sigma \left(\text{sim}(G_j^{Eins}, C_i^I) \right) \right], \quad (5)$$

where $\mathbf{T}^I = \frac{\text{sim}(I_i^{Eins}, C_j^G)}{\tau_{I-G}}$, $\mathbf{T}^G = \frac{\text{sim}(G_j^{Eins}, C_i^I)}{\tau_{I-G}}$, and τ_{I-G} denotes the temperature of instance-group discrimination.

To facilitate downstream inference of gene expression, we append a prediction head atop the fused image representations $\{I_i\}$. This module maps the learned self-supervised features to predicted gene expression profiles \hat{G}_i , and is optimized using a standard mean squared error (MSE) loss:

$$\mathcal{L}_{\text{pred}} = \frac{1}{N} \sum_{i=1}^N \left\| \hat{G}_i - G_i \right\|_2^2. \quad (6)$$

By combining the instance-level and instance-group losses with a task-specific prediction loss, Gene-DML has the ability to learn the alignments between WSIs and spatial gene expressions in terms of different semantic levels with the weight λ of instance-group loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{multi.ins}} + \lambda \mathcal{L}_{\text{cross}} + \mathcal{L}_{\text{pred}}. \quad (7)$$

4. Experiments and Results

4.1. Dataset and Implementation Details

In this study, we conducted experiments on three ST datasets that are collected in the HEST-1k [18] benchmark: HER2ST [2], STNet [14], and skinST [19]. Specifically, two breast cancer ST datasets – HER2ST, which focuses exclusively on HER2-positive breast tumor subtypes, and

Model	HER2ST			STNet			skinST		
	MSE↓	PCC(A)↑	PCC(H)↑	MSE↓	PCC(A)↑	PCC(H)↑	MSE↓	PCC(A)↑	PCC(H)↑
ST-Net [14] 20'	0.260 ± 0.04	0.194 ± 0.11	0.345 ± 0.16	0.209 ± 0.02	0.116 ± 0.06	0.223 ± 0.10	0.294 ± 0.07	0.274 ± 0.08	0.382 ± 0.08
HistoGene [25] 21'	0.314 ± 0.09	0.168 ± 0.12	0.302 ± 0.19	0.194 ± 0.05	0.100 ± 0.05	0.219 ± 0.12	0.270 ± 0.09	0.133 ± 0.06	0.261 ± 0.13
Hist2ST [45] 22'	0.285 ± 0.08	0.118 ± 0.10	0.248 ± 0.17	<u>0.181</u> ± 0.02	0.044 ± 0.02	0.099 ± 0.03	1.291 ± 0.65	0.004 ± 0.01	0.053 ± 0.01
EGN [42] 23'	0.241 ± 0.06	0.197 ± 0.11	0.328 ± 0.17	0.192 ± 0.02	0.111 ± 0.05	0.203 ± 0.09	0.281 ± 0.08	0.281 ± 0.06	0.388 ± 0.06
BLEEP [41] 23'	0.277 ± 0.05	0.151 ± 0.11	0.277 ± 0.16	0.235 ± 0.02	0.095 ± 0.05	0.193 ± 0.10	0.297 ± 0.05	0.269 ± 0.07	0.396 ± 0.08
TRIPLEX [7] 24'	<u>0.228</u> ± 0.07	<u>0.314</u> ± 0.14	<u>0.497</u> ± 0.17	0.202 ± 0.02	<u>0.206</u> ± 0.07	0.352 ± 0.10	<u>0.268</u> ± 0.09	<u>0.374</u> ± 0.07	<u>0.490</u> ± 0.07
M2OST [34] 25'	0.302 ± 0.06	0.231 ± 0.10	0.410 ± 0.17	0.278 ± 0.03	0.201 ± 0.07	<u>0.353</u> ± 0.10	0.271 ± 0.02	0.300 ± 0.05	<u>0.490</u> ± 0.05
Gene-DML (ours)	0.210 ± 0.03	0.331 ± 0.07	0.541 ± 0.08	0.179 ± 0.02	0.237 ± 0.09	0.384 ± 0.08	0.237 ± 0.04	0.433 ± 0.03	0.520 ± 0.07

Table 1. Comparison of cross-validation experiments with baselines. PCC(A) denotes the average PCC of all 250-genes, and PCC(H) denotes the average PCC of the top-50 (high) predictive genes. The **best results** are in bold; the second-best results are underlined.

STNet, which includes a broader range of breast cancer subtypes – provide approximately **13,620** and **68,050** spatially resolved spots, respectively. Additionally, skinST contains approximately **23,205** spots from spatial profiling of skin tissue sections. Each spot comprises its coordinates on the WSI sample and the corresponding gene expression profiles with around 20,000 genes for breast and skin cancer. To prevent data leakage, we ensure that samples from the same patient do not appear in both the training and testing datasets. Following [7], we adopt the one-patient-one-fold cross-validation (CV) approach for the HER2ST (8-fold) and skinST (4-fold) datasets; for STNet, which comprises 68 samples from 23 patients, we perform an 8-fold CV, ensuring that all samples from the same patient are assigned to the same fold. To evaluate the generalization performance of Gene-DML, we perform external tests on 3 breast Visium data and 2 skin Visium data. In our experiments, we use Pearson Correlation Coefficient (PCC), Mean Squared Error (MSE), and Mean Absolute Error (MAE) as evaluation metrics:

$$\text{PCC} = \frac{\sum_{i=1}^N (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^N (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^N (\hat{y}_i - \bar{\hat{y}})^2}}, \quad (8)$$

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2, \text{ and} \quad (9)$$

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|. \quad (10)$$

We report both the mean PCC for all genes (PCC(A)) and the mean PCC for highly predictive genes (PCC(H)). Specifically, PCC is computed for each gene across all spatial spots within a sample, and the highly predictive genes (HPG) are identified as the top-50 genes selected based on their average ranking across all CV folds.

At the data-preprocessing stage, we crop patches with 224×224 pixel patches centered on each spatial spot from WSIs to construct image tiles. For gene expression profiles, we select 250 genes per dataset (details refer to the supplementary), applying spot-wise normalization followed by

log transformation according to [14]. For feature grouping, we utilize K-means [44] method. Gene-DML is optimized using the Adam optimizer with an initial learning rate of $1e^{-4}$. The learning rate is decayed by a factor of 0.95 every 20 epochs using a StepLR scheduler. The batch size is set to 256 during training, and we report the mean of the model’s performance evaluated on all spots in every validation WSI during testing. All the experiments are trained on an NVIDIA RTX A6000 (48GB) GPU.

4.2. Cross-validation Performance of Gene-DML

To evaluate the performance of Gene-DML, we conducted cross-validation experiments on the public ST datasets, HER2ST, STNet, and skinST, and compared the spatial gene expression result predicted by Gene-DML with recent-year existing methods as baselines: ST-Net [14], EGN [42], BLEEP [41], HistoGene [25], Hist2ST [45], TRIPLEX [7], and M2OST [34]. We re-implement the baseline methods by their public code repositories under the same CV fold splitting settings on the same datasets.

In comparison with existing works, as shown in Tab. 1, Gene-DML consistently achieves the best performance across all metrics, MSE, PCC(A), and PCC(H), demonstrating its superiority in accurately capturing spatial gene expression patterns. Specifically, for HER2ST dataset, Gene-DML obtains the lowest MSE of 0.210, outperforming the second-best TRIPLEX (0.228) by 0.018. In terms of PCC(A), our method achieves 0.331, which surpasses TRIPLEX (0.314) by +0.017, and shows the most significant gain in PCC(H) at 0.541, improving by +0.044 over TRIPLEX (0.497). As for STNet dataset, Gene-DML achieves the best MSE of 0.179, a 0.002 reduction from Hist2ST (0.181), and attains a notable gain in PCC(A) with 0.237, improving upon TRIPLEX (0.206) by +0.031. In high-predictive gene correlation (PCC(H)), our method leads 0.384, ahead 0.353 of M2OST by +0.031. On the skin dataset skinST, Gene-DML achieves the most substantial improvements. Its MSE of 0.237 is better than 0.268 of TRIPLEX by +0.031. In PCC(A), it scores 0.433, significantly outperforming the second-best TRIPLEX (0.374) by +0.059. For PCC(H), Gene-DML leads with 0.520, an im-

Model	Breast Visium-1				Breast Visium-2				Breast Visium-3			
	MSE↓	MAE↓	PCC(A)↑	PCC(H)↑	MSE↓	MAE↓	PCC(A)↑	PCC(H)↑	MSE↓	MAE↓	PCC(A)↑	PCC(H)↑
ST-Net [14] 20'	0.423	0.505	-0.026	-0.000	0.395	0.492	0.091	0.193	0.424	0.508	-0.032	0.008
EGN [42] 23'	0.421	0.512	0.003	0.024	0.328	0.443	0.102	0.157	0.303	0.425	0.106	0.220
BLEEP [41] 23'	<u>0.367</u>	<u>0.470</u>	0.106	0.221	0.289	0.406	0.104	0.260	<u>0.298</u>	0.415	0.114	0.229
TRIPLEX [7] 24'	0.351	0.464	0.136	0.241	0.282	0.407	0.155	0.356	0.285	0.410	0.118	0.282
Gene-DML (ours)	0.410	0.512	0.160	0.415	<u>0.289</u>	<u>0.407</u>	0.192	0.411	0.301	0.408	0.155	0.393

Table 2. Comparison of generalization performance with baselines. The **best results** are in bold; the second-best results are underlined.

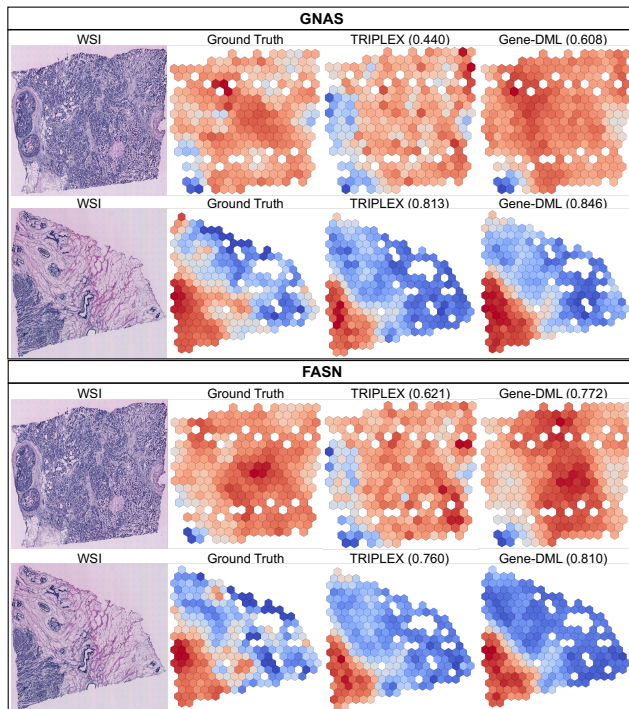


Figure 3. Visualization of cancer biomarker gene expression level prediction. All predicted values of gene expression are normalized to the range $[0, 1]$. The values in parentheses are the PCC between the ground truth of gene expression and the prediction on genes GNAS and FASN.

provement of $+0.030$ over M2OST and TRIPLEX (0.490). Due to space limitations, all MAE results are provided in the supplementary material.

Across all datasets and metrics, Gene-DML consistently sets a new performance benchmark, which reflects its capacity to capture biologically critical gene patterns. The consistent outperformance validates the strength of our dual-pathway multi-level alignments.

4.3. Generalization Performance of Gene-DML

Tab. 2 presents the generalization performance of Gene-DML compared to recent existing works across 3 unseen Visium ST datasets. The results are produced by using the pre-trained Gene-DML model on the breast cancer dataset (HER2ST). Our method achieves the highest Pearson Cor-

relation Coefficient (PCC) on both PCC(A) and PCC(H) genes across all datasets, highlighting its superior ability to capture biologically meaningful patterns that generalize beyond training distributions. Notably, Gene-DML achieves a new state-of-the-art PCC(A) of 0.160, 0.192, 0.155 and PCC(H) of 0.415, 0.411, 0.393 on breast Visium-1, -2, and -3, respectively, surpassing all prior approaches. While our MSE and MAE scores are not always the best, this can be attributed to Gene-DML’s emphasis on semantic alignment and structure-preserving embeddings rather than purely minimizing per-spot regression loss. The slight trade-off in MSE/MAE is acceptable, as PCC offers a more interpretable measure of functional correlation, especially for HPGs critical in downstream biological interpretation. The result of the additional 2 skin Visium data (NCBI463 and NCBI464) is included in the supplementary.

4.4. Visualization of Cancer Biomarker Genes

GNAS [1] and FASN [20] are two of the highly predictive genes across all datasets we utilized, and well-established cancer biomarker genes. Fig. 3 presents a qualitative comparison of gene expression prediction with indicating PCC values for GNAS and FASN across two representative tissue WSIs from various patients, with the original WSI, gene expression ground truth, TRIPLEX prediction, and Gene-DML prediction. Specifically, each hexagonal tile corresponds to a spatial transcriptomics (ST) spot, with its color indicating the predicted gene expression level, normalized to the range $[0, 1]$. Gene-DML consistently outperforms the model TRIPLEX with the second-best quantitative results in Tab. 1 in accurately recovering the spatial expression patterns, as evidenced by the visibly sharper regional expression and higher PCC values (e.g., 0.608 vs. 0.440 and 0.846 vs. 0.813 for GNAS; 0.772 vs. 0.621 and 0.810 vs. 0.760 for FASN) for both relatively hard (sample #1) and easy (sample #2) cases. The robust prediction of GNAS and FASN not only reflects the strength of features learned by the dual multi-scale discrimination but also is widely implicated in tumorigenesis and cancer progression, further underscoring the biological relevance and clinical utility of the Gene-DML framework for spatial biomarker discovery. Visualizations of additional samples are provided in the supplementary material.

Multi-scale I-L Dis.	Global Rep.	Cross-level I-G Dis.	Feature Grouping	MSE↓	MAE↓	PCC(A)↑	PCC(H)↑
✗	✗	✗	✗	0.297	0.430	0.269	0.396
✓	✓	✗	✗	0.282	0.414	0.308	0.449
✗	✗	✓	✓	0.280	0.413	0.320	0.478
✓	✗	✓	✓	0.247	0.394	0.421	0.478
✓	✓	✓	✗	0.253	0.399	0.398	0.456
✓	✓	✓	✓	0.237	0.350	0.433	0.520

Table 3. Ablation study on framework components on the dataset skinST. Abbreviations in this table include instance-level discrimination (I-L Dis.), global-scale representation (Global Rep.), and instance-group discrimination (I-G Dis.).

k	MSE↓	MAE↓	PCC(A)↑	PCC(H)↑
15	0.252	0.363	0.426	0.516
20	0.250	0.363	0.433	0.503
25	0.237	0.350	0.433	0.520
30	0.240	0.352	0.429	0.513

Table 4. Ablation study on the number of k -groups in the feature grouping process on the dataset skinST.

4.5. Ablation Study

We performed extensive experiments to assess the effectiveness of each framework module with its key component, values of hyperparameters k (feature group size), λ (the weight of instance-group alignment loss), and τ_{I-G} (the temperature of instance-group similarity calculation).

We report the ablation results on each key component of Gene-DML, as illustrated in Tab. 3. This underscores the effectiveness of the **multi-scale instance-level discrimination** module and the **cross-level instance-group discrimination** module, together with their corresponding *key components*: the **global-scale representation** and **feature grouping**. Incorporating the multi-scale instance-level discrimination improves performance 0.015 on MSE, +0.039 on PCC(A), and +0.053 on PCC(H). Similarly, adding cross-level instance-group discrimination yields a gain 0.017 on MSE, 0.051 on PCC(A), and 0.082 on PCC(H) from the plain bimodal framework. When both modules are integrated, Gene-DML achieves the best performance across all metrics, showing a cumulative improvement of 0.060 MSE, +0.164 PCC(A), and +0.124 PCC(H) over the baseline. Besides, comparative experiments with and without the global-scale representation and feature grouping further highlight their essential contribution to the effectiveness of the corresponding modules. These results validate the complementary benefits of a dual-pathway multi-level discrimination strategy.

Tab. 4 demonstrates the impact of varying the number of groups k in the feature grouping (Sec. 3.2.1). The model performance is relatively robust across different values but achieves the best results when $k = 25$, with

λ	MSE↓	MAE↓	PCC(A)↑	PCC(H)↑
0.3	0.263	0.380	0.391	0.430
0.5	0.259	0.373	0.402	0.493
0.8	0.237	0.350	0.433	0.520
1	0.241	0.365	0.426	0.510

Table 5. Ablation study on the weight of instance-group loss λ on the dataset skinST. $\lambda = 1$ indicates that the weight of instance-group loss is not applied in the experimental setting.

τ_{I-G}	MSE↓	MAE↓	PCC(A)↑	PCC(H)↑
0.05	0.269	0.380	0.420	0.513
0.07	0.237	0.350	0.433	0.520
0.1	0.273	0.381	0.412	0.490

Table 6. Ablation study on the temperature of instance-group similarity calculation τ_{I-G} on the dataset skinST.

the lowest MSE and MAE, and the highest PCC(A) and PCC(H). These results suggest that an appropriately chosen group size provides a good balance between semantic expressiveness and discrimination, facilitating more effective instance-group alignment. Tab. 5 presents an ablation study on the impact of the weighting coefficient λ for the instance-group contrastive loss. The results show that incorporating instance-group discrimination consistently improves performance across all metrics. Specifically, when $\lambda = 0.8$, the best overall results are obtained, achieving the lowest MSE and MAE, along with the highest PCC scores. This indicates that moderately weighting the instance-group loss effectively enhances cross-modal instance-group alignment without overwhelming the training objective. Tab. 6 shows the effect of the temperature parameter τ_{I-G} used in the instance-group similarity calculation. τ_{I-G} controls the sharpness of the softmax distribution over similarity scores. The results show that setting $\tau_{I-G} = 0.07$ achieves optimal performance across all evaluation metrics. The ablation of HER2ST and STNet is included in the supplementary.

5. Conclusion

In this work, we introduced Gene-DML, a novel framework that advances gene expression prediction from histopathology images by modeling dual-pathway multi-level discrimination. By jointly integrating multi-scale instance-level and cross-level instance-group discriminations, Gene-DML effectively captures both fine-grained and high-level semantic correspondences across visual and transcriptomic modalities. Validated through extensive experiments, our dual-pathway discrimination design facilitates robust and generalizable cross-modal representation learning in spatial transcriptomic data, enabling the model to better reflect the biological and spatial complexities inherent in tissue samples.

References

- [1] Hafeez Abiola Afolabi, Salzihan Md Salleh, Zaidi Zakaria, Ewe Seng Ch'ng, Siti Norasikin Mohd Nafi, Ahmad Aizat Bin Abdul Aziz, Ahmad Adebayo Irekeola, Yusuf Wada, and Sameer Badri Al-Mhanna. A gnas gene mutation's independent expression in the growth of colorectal cancer: A systematic review and meta-analysis. *Cancers*, 14(22), 2022. 7
- [2] Alma Andersson, Ludvig Larsson, Linnea Stenbeck, Fredrik Salmén, Anna Ehinger, Sunny Z Wu, Ghamdan Al-Eryani, Daniel Roden, Alex Swarbrick, Åke Borg, et al. Spatial deconvolution of her2-positive breast cancer delineates tumor-associated cell type interactions. *Nature communications*, 12(1):6012, 2021. 5
- [3] Haodong Chen, Runnan Chen, Qiang Qu, Zhaoqing Wang, Tongliang Liu, Xiaoming Chen, and Yuk Ying Chung. Beyond gaussians: Fast and high-fidelity 3d splatting with linear kernels. *arXiv preprint arXiv:2411.12440*, 2024. 3
- [4] Kok Hao Chen, Alistair N Boettiger, Jeffrey R Moffitt, Siyuan Wang, and Xiaowei Zhuang. Spatially resolved, highly multiplexed rna profiling in single cells. *Science*, 348(6233):aaa6090, 2015. 1
- [5] Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16144–16155, 2022. 2
- [6] Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Bowen Chen, Andrew Zhang, Daniel Shao, Andrew H Song, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 2024. 3
- [7] Youngmin Chung, Ji Hun Ha, Kyeong Chan Im, and Joo Sang Lee. Accurate spatial gene expression prediction by integrating multi-resolution features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11591–11600, 2024. 1, 2, 3, 4, 6, 7
- [8] Saisai Ding, Juncheng Li, Jun Wang, Shihui Ying, and Jun Shi. Multi-scale efficient graph-transformer for whole slide image classification. *IEEE Journal of Biomedical and Health Informatics*, 27(12):5926–5936, 2023. 2
- [9] Chee-Huat Linus Eng, Michael Lawson, Qian Zhu, Ruben Dries, Noushin Koulana, Yodai Takei, Jina Yun, Christopher Cronin, Christoph Karp, Guo-Cheng Yuan, et al. Transcriptome-scale super-resolved imaging in tissues by rna seqfish+. *Nature*, 568(7751):235–239, 2019. 1
- [10] Jianan Fan, Dongnan Liu, Hang Chang, and Weidong Cai. Learning to generalize over subpartitions for heterogeneity-aware domain adaptive nuclei segmentation. *International Journal of Computer Vision*, 2024. 2
- [11] Jianan Fan, Dongnan Liu, Canran Li, Hang Chang, Heng Huang, Filip Braet, Mei Chen, and Weidong Cai. Revisiting adaptive cellular recognition under domain shifts: A contextual correspondence view. In *Computer Vision – ECCV 2024*. Springer Nature Switzerland, 2024. 2
- [12] Bo Fang, Jianan Fan, Dongnan Liu, Hang Chang, Gerald J. Shami, Filip Braet, and Weidong Cai. Scsam: Debiasing morphology and distributional variability in subcellular semantic segmentation. *arXiv preprint arXiv:2507.17149*, 2025. 2
- [13] Aniruddha Ganguly, Debolina Chatterjee, Wentao Huang, Jie Zhang, Alisa Yurovsky, Travis Steele Johnson, and Chao Chen. Merge: Multi-faceted hierarchical graph-based gnn for gene expression prediction from whole slide histopathology images. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15611–15620, 2025. 1
- [14] Bryan He, Ludvig Bergensträhle, Linnea Stenbeck, Abubakar Abid, Alma Andersson, Åke Borg, Jonas Maaskola, Joakim Lundeberg, and James Zou. Integrating spatial gene expression and breast tumour morphology via deep learning. *Nature biomedical engineering*, 4(8): 827–834, 2020. 1, 2, 5, 6, 7
- [15] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 4
- [16] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 2
- [17] Wentao Huang, Meilong Xu, Xiaoling Hu, Shahira Abousamra, Aniruddha Ganguly, Saarthak Kapse, Alisa Yurovsky, Prateek Prasanna, Tahsin Kurc, Joel Saltz, et al. Rankbygene: Gene-guided histopathology representation learning through cross-modal ranking consistency. *arXiv preprint arXiv:2411.15076*, 2024. 1, 2
- [18] Guillaume Jaume, Paul Doucet, Andrew Song, Ming Yang Lu, Cristina Almagro Pérez, Sophia Wagner, Anurag Vaidya, Richard Chen, Drew Williamson, Ahrong Kim, et al. Hest-1k: A dataset for spatial transcriptomics and histology image analysis. *Advances in Neural Information Processing Systems*, 37:53798–53833, 2024. 5
- [19] Andrew L Ji, Adam J Rubin, Kim Thrane, Sizun Jiang, David L Reynolds, Robin M Meyers, Margaret G Guo, Benson M George, Annelie Mollbrink, Joseph Bergensträhle, et al. Multimodal analysis of composition and spatial architecture in human squamous cell carcinoma. *Cell*, 182(2): 497–514, 2020. 5
- [20] Weihua Jiang, Xiao-Liang Xing, Chenguang Zhang, Lina Yi, Wenting Xu, Jianghua Ou, and Ning Zhu. Met and fasn as prognostic biomarkers of triple negative breast cancer: A systematic evidence landscape of clinical study. *Frontiers in Oncology*, 11:604801, 2021. 7
- [21] Yunfan Li, Mouxiang Yang, Dezhong Peng, Taihao Li, Jiantao Huang, and Xi Peng. Twin contrastive learning for online clustering. *International Journal of Computer Vision*, 130(9):2205–2221, 2022. 3
- [22] Yuxiang Lin, Ling Luo, Ying Chen, Xushi Zhang, Zihui Wang, Wenxian Yang, Mengsha Tong, and Rongshan Yu. St-align: A multimodal foundation model for image-gene alignment in spatial transcriptomics. *arXiv preprint arXiv:2411.16793*, 2024. 1, 2
- [23] Gabriel Mejia, Daniela Ruiz, Paula Cárdenas, Leonardo Manrique, Daniela Vega, and Pablo Arbeláez. Enhancing

- gene expression prediction from histology images with spatial transcriptomics completion. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 91–101. Springer, 2024. 1
- [24] Wenwen Min, Zhiceng Shi, Jun Zhang, Jun Wan, and Changmiao Wang. Multimodal contrastive learning for spatial gene expression prediction using histology images. *Briefings in Bioinformatics*, 25(6):bbae551, 2024. 1, 2
- [25] Minxing Pang, Kenong Su, and Mingyao Li. Leveraging information in spatial transcriptomics to predict super-resolution gene expression from histology images in tumors. *BioRxiv*, pages 2021–11, 2021. 1, 2, 6
- [26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 3, 5
- [27] Anjali Rao, Dalia Barkley, Gustavo S França, and Itai Yanai. Exploring tissue architecture using spatial transcriptomics. *Nature*, 596(7871):211–220, 2021. 1
- [28] Yang Song, Hang Chang, Yang Gao, Sidong Liu, Donghao Zhang, Junen Yao, Wojciech Chrzanowski, and Weidong Cai. Feature learning with component selective encoding for histopathology image classification. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 257–260, 2018. 2
- [29] Patrik L Ståhl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakob O Westholm, Mikael Huss, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82, 2016. 1
- [30] Aiham Taleb, Matthias Kirchler, Remo Monti, and Christoph Lippert. Contig: Self-supervised multimodal contrastive learning for medical imaging with genetics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20908–20921, 2022. 3
- [31] Asher Trockman and J Zico Kolter. Patches are all you need? *arXiv preprint arXiv:2201.09792*, 2022. 2
- [32] Zhenchen Wan, Yanwu Xu, Zhaoqing Wang, Feng Liu, Tongliang Liu, and Mingming Gong. Ted-viton: Transformer-empowered diffusion models for virtual try-on. *arXiv preprint arXiv:2411.17017*, 2024. 2
- [33] Hao Wang, Euijoon Ahn, and Jinman Kim. A dual-branch self-supervised representation learning framework for tumor segmentation in whole slide images. *arXiv preprint arXiv:2303.11019*, 2023. 2
- [34] Hongyi Wang, Xiuju Du, Jing Liu, Shuyi Ouyang, Yen-Wei Chen, and Lanfen Lin. M2ost: Many-to-one regression for predicting spatial transcriptomics from digital pathology images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7709–7717, 2025. 2, 6
- [35] Tianyi Wang, Jianan Fan, Dingxin Zhang, Dongnan Liu, Yong Xia, Heng Huang, and Weidong Cai. MIRROR: Multi-Modal Pathological Self-Supervised Representation Learning via Modality Alignment and Retention. *IEEE Transactions on Medical Imaging*, 2025. 1
- [36] Xiao Wang, William E Allen, Matthew A Wright, Emily L Sylwestrak, Nikolay Samusik, Sam Vesuna, Kathryn Evans, Cindy Liu, Charu Ramakrishnan, Jia Liu, et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science*, 361(6400):eaat5691, 2018. 1
- [37] Xudong Wang, Ziwei Liu, and Stella X Yu. Unsupervised feature learning by cross-level instance-group discrimination. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12586–12595, 2021. 2
- [38] Xiaofei Wang, Xingxu Huang, Stephen Price, and Chao Li. Cross-modal diffusion modelling for super-resolved spatial transcriptomics. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 98–108. Springer, 2024. 1
- [39] Zhaoqing Wang, Ziyu Chen, Yaqian Li, Yandong Guo, Jun Yu, Mingming Gong, and Tongliang Liu. Mosaic representation learning for self-supervised visual pre-training. In *The eleventh international conference on learning representations*, 2023. 3
- [40] Tiange Xiang, Yang Song, Chaoyi Zhang, Dongnan Liu, Mei Chen, Fan Zhang, Heng Huang, Lauren O’Donnell, and Weidong Cai. Dsnet: A dual-stream framework for weakly-supervised gigapixel pathology image analysis. *IEEE Transactions on Medical Imaging*, pages 2180–2190, 2022. 2
- [41] Ronald Xie, Kuan Pang, Sai Chung, Catia Perciani, Sonya MacParland, Bo Wang, and Gary Bader. Spatially resolved gene expression prediction from histology images via bimodal contrastive learning. *Advances in Neural Information Processing Systems*, 36:70626–70637, 2023. 1, 2, 5, 6, 7
- [42] Yan Yang, Md Zakir Hossain, Eric A Stone, and Shafin Rahman. Exemplar guided deep neural network for spatial transcriptomics analysis of gene expression prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5039–5048, 2023. 1, 2, 6, 7
- [43] Yan Yang, Md Zakir Hossain, Xuesong Li, Shafin Rahman, and Eric Stone. Spatial transcriptomics analysis of zero-shot gene expression prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 492–502. Springer, 2024. 1
- [44] Jieping Ye, Zheng Zhao, and Mingrui Wu. Discriminative k-means for clustering. *Advances in neural information processing systems*, 20, 2007. 4, 6
- [45] Yuansong Zeng, Zhuoyi Wei, Weijiang Yu, Rui Yin, Yuchen Yuan, Bingling Li, Zhonghui Tang, Yutong Lu, and Yuedong Yang. Spatial transcriptomics prediction from histology jointly through transformer and graph neural networks. *Briefings in Bioinformatics*, 23(5):bbac297, 2022. 1, 2, 6
- [46] Sichen Zhu, Yuchen Zhu, Molei Tao, and Peng Qiu. Diffusion generative modeling for spatially resolved gene expression inference from histology images. *arXiv preprint arXiv:2501.15598*, 2025. 2
- [47] Mohammadreza Zolfaghari, Yi Zhu, Peter Gehler, and Thomas Brox. Crossclr: Cross-modal contrastive learning for multi-modal video representations. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1450–1459, 2021. 3