COAST: Context-Aware Differential Learning for Gene Expression Prediction in Spatial Transcriptomics

Keunho Byeon[1], Sunhong Park[1], Jeewoo Lim[1], and Jin Tae Kwak[1]

School of Electrical Engineering, Korea University, Seoul 02841, Republic of Korea
{bkh5922,sunhongpapa,jeewoolim,jkwak}@korea.ac.kr

Abstract. Spatial transcriptomics enables profiling of spatial gene expression but is limited by high cost and low throughput, motivating prediction from H&E histopathology images. Existing context-aware methods mainly supervise absolute expression, while relative expression relationships between spots are rarely used explicitly. We propose COAST, a context-aware differential learning framework for spatial gene expression prediction. COAST conditions the local and global context features with type-specific modulation and aggregates the target and context spot tokens using a Transformer encoder to capture both fine-grained local patterns and slide-level structure. It is trained with a joint objective that combines absolute expression regression with signed differential regression between the target and context spots. Experiments on multiple spatial transcriptomics datasets show consistent improvements in correlation- and distribution-based metrics, demonstrating the effectiveness of context-aware differential learning for histology-based spatial gene expression prediction.

Keywords: Spatial Transcriptomics · Differential Learning · Context-Aware

1 Introduction

Spatial transcriptomics (ST) has emerged as a transformative technology that bridges morphological observations with molecular profiling by providing gene expression measurements within their native spatial context. Despite its great potential, current ST platforms, such as 10x Genomics Visium and Slide-seq, are constrained by high experimental costs, limited resolution, and relatively low throughput. These practical barriers severely restrict their widespread adoption in routine clinical workflows. Consequently, there is growing interest in developing and deploying computational approaches capable of inferring spatial gene expression directly from universally available and cost-effective hematoxylin and eosin (H&E) stained histopathology images. Inferring gene expression from histology is inherently challenging due to the complex and highly non-linear relationship between tissue morphology and transcriptomic states. Early approaches framed this as an independent patch-level regression problem, focusing on local

feature representations extracted from individual image patches and regressing expression values at each spot in isolation. However, gene expression is fundamentally shaped by spatial context. Morphologically similar regions can exhibit distinct expression levels depending on their specific microenvironment, structural compartment, and global tissue composition. To address this, recent methods introduced neighbor aggregation, graph-based propagation, or attention-based mechanisms to incorporate spatial context from adjacent regions.

Despite these advancements, two critical limitations persist. First, most existing models are optimized to regress the absolute expression magnitude at each spot, making predictions sensitive to slide-level variability and protocol-dependent shifts. Second, although spatial context is often provided as input, models are rarely supervised to explicitly preserve relative spatial relationships. In biological tissues, spatial gradients and relative changes between regions (such as tumor-stroma boundaries) frequently encode more stable and structurally meaningful signals than absolute counts.

To address these limitations, we propose **COAST**, a context-aware differential learning framework for spatial gene expression prediction from H&E histopathology images. COAST seamlessly integrates heterogeneous contextual information through type-specific contextual feature modulation, enabling adaptive conditioning of local and global context features. Critically, COAST is optimized under a joint objective that combines standard absolute expression regression with signed differential prediction between a target spot and its contextual spots, thereby directly supervising spatial relationships. Experiments across multiple public ST datasets demonstrate consistent improvements over existing baselines. Furthermore, the reconstructed gene expression representations retain clinically meaningful prognostic information in downstream survival prediction, highlighting the practical relevance of context-aware differential learning.

2 Related Works

2.1 Histology-based Gene Expression Prediction

Several studies have investigated predicting spatial gene expression directly from H&E histopathology images. Early approaches like STNet [7] introduced convolutional encoders to predict spot-level gene expression from local image patches. Subsequent methods sought to capture broader tissue architecture by integrating spatial positional encoding and Transformers (Hist2ST [24]), constructing spatial graphs over spots for neighborhood-aware prediction (TCGN [22]), and employing retrieval-based transfer modeling with graph neural networks (EGNv2 [23]). More recently, NH2ST [17] introduced a dual-branch architecture with cross-attention and contrastive learning, while PEKA [14] leveraged a single-cell foundation model for parameter-efficient knowledge transfer. TRIPLEX [3] fuses separately encoded target, neighboring, and slide-level global views. M2OST [20] jointly processes multiple WSI pyramid levels using a many-to-one Transformer. MERGE [6] models local and long-range spot dependencies using a hierarchical

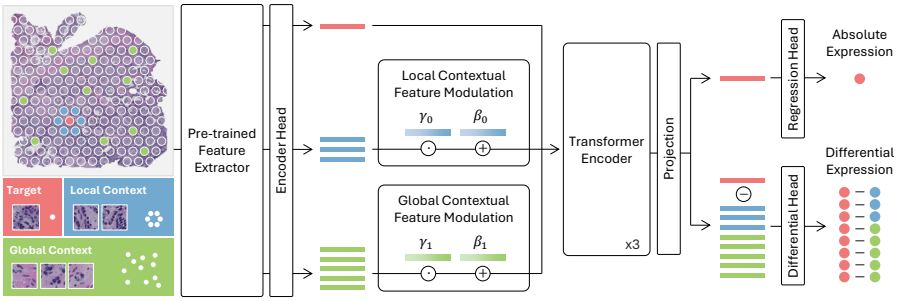

Fig. 1. Overview of the COAST framework. COAST integrates local and global context spots through contextual feature modulation and Transformer-based aggregation. The architecture jointly optimizes absolute and differential regression.

GNN constructed from spatial and feature-space clusters. Although these methods progressively incorporated spatial context, their learning objectives primarily remain focused on absolute spot-wise expression regression.

2.2 Relative and Differential Learning Objectives

Relative and differential supervision has become a cornerstone of robust representation learning. Methods such as Lifted Structured Embedding [13] and Ranked List Loss [21] enforce relational constraints among samples. DIOR-ViT [9] introduced a differential ordinal regression objective that explicitly models ordered relationships between samples to improve structural consistency and robustness. However, explicitly supervising signed gene-wise expression differences between spatially related regions remains underexplored in histology-based ST prediction.

2.3 Feature-wise Affine Modulation

Feature-wise affine modulation has been widely adopted for conditioning intermediate representations. Representative methods include AdaIN [8], FiLM [16], and SPADE [15], which apply learned scaling and shifting parameters to adapt features based on external signals.

3 Methods

The COAST framework is designed to learn the gene expression patterns at a target spot through explicit interactions with its heterogeneous spatial context (Fig. 1). The framework consists of four major components: (1) a feature extractor, (2) context-specific feature modulation, (3) spatio-relational Transformer, and (4) absolute and differential prediction heads.

3.1 Feature Extraction and Context Sampling

Let a whole slide image (WSI) be defined as $I = \{p_i\}_{i=1}^{N}$, where p_i is the image patch associated with spot i and N is the total number of spots. For a given target spot t, let $x_t \in \mathbb{R}^D$ denote its feature vector, extracted using a pre-trained feature extractor. To effectively capture both the tissue microenvironment and global structure, we sample a context feature set $\mathcal{C}_t = \{x_1, \ldots, x_K\}$ containing K spot features from I. This context set consists of two distinct subsets: $\mathcal{C}_t = \mathcal{C}_{\mathrm{L}} \cup \mathcal{C}_{\mathrm{G}}$, where $\mathcal{C}_{\mathrm{L}}$ contains K_{L} local context spot features sampled via k-nearest neighbors, $\mathcal{C}_{\mathrm{G}}$ comprises K_{G} global context spot features, and $K = K_{\mathrm{L}} + K_{\mathrm{G}}$. To construct $\mathcal{C}_{\mathrm{G}}$ during training, we randomly sample spot features from I to enhance robustness and diversity. During inference, we employ a deterministic stratified grid strategy to ensure reproducible and representative context selection. Finally, each context spot $x_k \in \mathcal{C}_t$ is associated with a type indicator $c_k \in \{0, 1\}$, where 0 and 1 represent local and global types, respectively.

3.2 Context-Specific Modulation and Spatio-Relational Transformer

Both target and context spot features are projected into a latent space using a shared encoder head Φ_{head}, which consists of a linear layer followed by a LeakyReLU activation and dropout:

$$h_t = \Phi_{head}(x_t), \quad h_k = \Phi_{head}(x_k) \tag{1}$$

where $h_t, h_k \in \mathbb{R}^d$. Although context spot features are sampled from spatially proximal (local) and distant (global) regions, the individual features x_k only carry localized visual information, without retaining their spatial context. Accordingly, naively aggregating local and global context with the target spot can conflate heterogeneous signals and weaken type-specific contributions during context integration. To effectively integrate these heterogeneous context sources and resolve this spatial ambiguity, we employ a contextual feature modulation mechanism with two sets of scale (γ) and shift (β) parameters: $\gamma_0, \beta_0 \in \mathbb{R}^d$ for local context ($c_k = 0$) and $\gamma_1, \beta_1 \in \mathbb{R}^d$ for global context ($c_k = 1$). We learn these context-specific scale and shift parameters via embedding layers to modulate the context features:

$$\tilde{h}_k = h_k \odot (1 + \gamma_{c_k}) + \beta_{c_k} \tag{2}$$

where $\odot$ denotes element-wise multiplication. This allows the model to adaptively condition the latent representations based on their spatial context type.

We then form a token sequence by concatenating the target token and the modulated context tokens, $[h_t; \tilde{h}_1, \ldots, \tilde{h}_K]$, and feed it into a standard L-layer Transformer encoder to aggregate heterogeneous context. The Transformer encoder outputs refined context-aware representations $Z = [z_t; z_1, \ldots, z_K]$.

3.3 Absolute and Differential Prediction Heads

We project the refined features Z using a shared projection head Φ_{proj} to obtain the final spot representation $f_t = \Phi_{proj}(z_t) \in \mathbb{R}^D$ and context representations

Table 1. Comparison of gene expression prediction performance across different models. Best results are shown in **bold**.

Model	PCC ↑	MI ↑	AUC_{0vNZ} ↑	AUC_{Q50} ↑	NRMSE ↓
STNet	0.0091 ± 0.01	0.0192 ± 0.01	0.4784 ± 0.04	0.5024 ± 0.01	0.1829 ± 0.04
Hist2ST	0.0516 ± 0.00	0.0788 ± 0.03	0.4933 ± 0.02	0.5000 ± 0.00	0.1528 ± 0.03
TCGN	0.1920 ± 0.09	0.0628 ± 0.02	0.5571 ± 0.07	0.6066 ± 0.05	0.1961 ± 0.06
EGNv2	0.2469 ± 0.08	0.0739 ± 0.03	0.6132 ± 0.04	0.6453 ± 0.05	0.1480 ± 0.04
NH2ST	0.3279 ± 0.07	0.1011 ± 0.02	0.6149 ± 0.07	0.6633 ± 0.03	0.1449 ± 0.04
PEKA	0.3863 ± 0.08	0.1109 ± 0.03	0.6508 ± 0.04	0.6879 ± 0.03	0.1322 ± 0.02
COAST	$\mathbf{0.4262 \pm 0.08}$	$\mathbf{0.1487 \pm 0.03}$	$\mathbf{0.6567 \pm 0.06}$	$\mathbf{0.7170 \pm 0.02}$	$\mathbf{0.1245 \pm 0.02}$

$f_k = \Phi_{proj}(z_k) \in \mathbb{R}^D$. The projection head Φ_{proj} includes a sequence of LN, a linear layer, LeakyReLU, and dropout. To capture both absolute and relative expression patterns, we introduce two independent linear prediction heads: (1) Regression Head (ϕ_{reg}): Predicts the absolute gene expression vector $\hat{y}_t = \phi_{reg}(f_t)$ and (2) Differential Head (ϕ_{diff}): Predicts the expression difference between the target and each context spot from their feature difference, $\Delta\hat{y}_{t,k} = \phi_{diff}(f_t - f_k)$, for $k = 1, \cdots, K$.

3.4 Training Objectives

The total objective function $\mathcal{L}$ is defined as the weighted sum of the absolute regression loss $\mathcal{L}_{reg}$ and the differential regression loss $\mathcal{L}_{diff}$:

$$\mathcal{L} = \underbrace{\|y_t - \hat{y}_t\|^2}_{\mathcal{L}_{reg}} + w_{\text{diff}} \frac{1}{K} \sum_{k=1}^{K} \underbrace{\|\Delta y_{t,k} - \Delta\hat{y}_{t,k}\|^2}_{\mathcal{L}_{diff}} \tag{3}$$

where $\Delta y_{t,k} = y_t - y_k$ and y_t and y_k are ground-truth expression vectors for the target and context spots, and w_{diff} is a hyperparameter. This explicit supervision of relative differences encourages the model to learn stable spatial gradients of gene expression, enhancing robustness against slide-level batch effects.

4 Experiments

4.1 Implementation Details

We employed UNI [2] as a frozen pre-trained feature extractor. COAST used an input hidden dimension D=1024, Transformer L=3 layers, embedding dimension d=512 and H=8 attention heads. We set $K_{\text{L}} = 18$ and $K_{\text{G}} = 72$ ($K = 90$ total) and set the differential weight to $w_{\text{diff}} = 1.0$. The model was optimized with Adam (learning rate 2×10^{-4}, weight decay 1×10^{-4}) for 10 epochs with batch size 1, selecting the best checkpoint on validation PCC. All experiments were conducted on a single NVIDIA H200 GPU using Python 3.12 and PyTorch 2.7.

4.2 Datasets

SpaRED [11] provides multiple spatial transcriptomics cohorts collected from diverse tissues and species. In this study, we used seven SpaRED datasets that include predefined test splits and contain 128 target genes: Abalo Human Squamous Cell Carcinoma (AHSCC) [1], Erickson Human Prostate Cancer P1 (EHPCP1) [5], Mirzazadeh Mouse Bone (MMBO) [12], Mirzazadeh Mouse Brain P1 (MMBP1) and P2 (MMBP2) [12], Vicari Mouse Brain (VMB) [18], and Villacampa Lung Organoid (VLO) [19]. The target genes were dataset-specific, Moran's I-based highly variable genes provided by SpaRED. For each spot, we directly used the provided 224×224 H&E patch and the Transcripts Per Million (TPM) normalized, log1p-transformed expression values released in SpaRED.

4.3 Comparison Results

We compared COAST with six representative and competitive baseline models, including STNet, Hist2ST, TCGN, EGNv2, NH2ST, and PEKA. For baseline models, we used the official implementations, retained their original image encoders to preserve the intended architectures, and followed the reported default hyperparameter settings while matching the target genes, data splits, and evaluation metrics. Table 1 summarizes the average performance across seven SpaRED datasets. We report gene-wise averages over $G = 128$ genes using Pearson Correlation Coefficient (PCC), Mutual Information (MI) estimated with k-nearest neighbors ($k = 5$), ROC-$\text{AUC}_{0\text{vNZ}}$ for distinguishing zero from non-zero expression, ROC-AUC_{Q50} for discriminating values above and below the median, and Normalized Root Mean Square Error (NRMSE). COAST achieved the best performance on all five evaluation metrics, obtaining substantial performance gains over all baselines: $0.0399 \sim 0.4171$ in PCC, $0.0378 \sim 0.1295$ in MI, $0.0059 \sim 0.1783$ in $\text{AUC}_{0\text{vNZ}}$, $0.0291 \sim 0.2170$ in AUC_{Q50}, and $0.0077 \sim 0.0716$ in NRMSE. When PCC was computed across biologically meaningful gene categories, COAST outperformed PEKA and NH2ST for tumor-associated genes (0.4100 vs. 0.3959 and 0.3096), housekeeping/basal genes (0.4100 vs. 0.3896 and 0.3261), hypoxia/metabolism genes (0.3414 vs. 0.3088 and 0.2518), immune/TME genes (0.3436 vs. 0.3239 and 0.2581), and stromal/ECM or tumor–stroma boundary genes (0.3630 vs. 0.3153 and 0.2490), respectively. These results indicate that the improvements of COAST extend across multiple biologically meaningful gene groups. We further analyzed per-dataset performance using radar charts and Top-1 frequency rates, measuring how often a method achieves the best result among all competing models (Fig. 2). COAST consistently formed the largest polygons in radar charts and obtained the highest Top-1 frequency, indicating superior and balanced performance across diverse datasets and evaluation metrics. Qualitative assessment via Leiden clustering on reconstructed expression (Fig. 3) further reveals that COAST captures more coherent spatial domains, reflecting improved structural consistency.

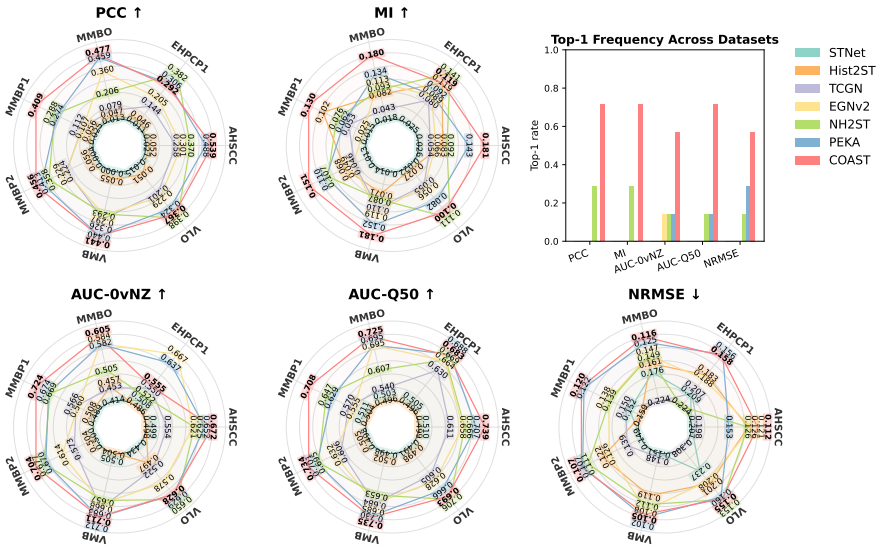

Fig. 2. Comparison of per-dataset gene expression prediction performance using radar charts and Top-1 frequency rates.

Table 2. Ablation study on context sampling, differential objective weight, and contextual feature modulation strategy.

K_{L}	K_{G}	PCC ↑	w_{diff}	PCC ↑	Modulation	PCC ↑
18	36	0.4178 ± 0.07	0.0	0.3785 ± 0.09	w/o	0.3999 ± 0.08
18	108	0.4214 ± 0.07	0.5	0.3762 ± 0.09	γ	0.4128 ± 0.07
36	72	0.3926 ± 0.13	2.0	0.3914 ± 0.12	β	0.4131 ± 0.07
18	72	0.4262 ± 0.08	1.0	0.4262 ± 0.08	γ and β	0.4262 ± 0.08

4.4 Ablation Study

Table 2 demonstrates ablation experiments using PCC to validate three key components of COAST: (1) context sampling size (K_{L} and K_{G}), (2) a differential objective weight w_{diff}, and (3) context-specific modulation. For context sampling, the combination K_{L}=18 and K_{G}=72 yielded the best performance (PCC=0.4262), whereas increasing either K_{L} or K_{G} degraded performance. The ablation of w_{diff} revealed the critical role of differential learning. Removing the differential objective ($w_{\mathrm{diff}} = 0$) led to a substantial decrease in PCC by 0.0477. While optimal performance was achieved at w_{diff}=1.0, increasing the weight to 2.0 resulted in a performance drop. Regarding context-specific modulation, jointly applying scaling and shifting (γ and β) achieved the strongest performance. Overall, these results suggest that COAST's gains are not driven by a single component but arise from complementary design choices. Across these three experiments, most ablation variants remained superior to PEKA (the strongest baseline), further indicating the superiority of COAST.

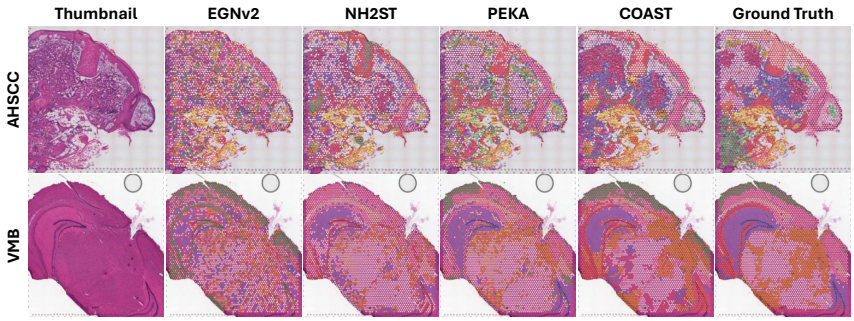

Fig. 3. Visualization of Leiden clustering on predicted and ground-truth expression.

4.5 Downstream Task: Overall Survival Prediction on TCGA-LUAD

To evaluate whether the gene expression representations reconstructed by COAST preserve clinically relevant prognostic signals, we conducted overall survival (OS) prediction on the TCGA-LUAD cohort. Patient-level data splits were used with 5-fold cross-validation. Patch-level image features were extracted using the frozen UNI [2], and gene expression representations were inferred from a model pre-trained on the VLO dataset without fine-tuning. We employed CLAM-SB [10] for slide-level risk prediction and optimized the model using the Cox proportional hazards partial likelihood [4].

We compared three input configurations: (1) image features only (Img-Only), (2) image features combined with PEKA-predicted expression representations (Img+PEKA), and (3) image features combined with COAST-predicted expression representations (Img+COAST). For a biological reference, a bulk RNA baseline was trained using a standard Cox proportional hazards model directly on patient-level bulk transcriptomic profiles without image features. OS prediction performance was quantified using the concordance index (C-index). As shown in Table 3, Img+COAST achieved an average C-index of 0.6029, comparable to bulk RNA (0.6042) and outperforming Img-Only (0.5845) and Img+PEKA (0.5904). These results provide evidence that COAST-derived representations retain prognostically relevant signals from morphology alone.

5 Conclusion

In this paper, we present COAST, a context-aware differential learning framework for predicting spatial gene expression from H&E-stained histopathology images. By incorporating both local microenvironment and global tissue structure as heterogeneous context spots, COAST effectively overcomes the limitations of existing methods that primarily focused on patch-level regression. Across seven datasets, COAST consistently outperforms representative baselines. Beyond prediction accuracy, COAST-derived expression representations preserve clinically meaningful prognostic signals, improving OS prediction on TCGA-LUAD when

Table 3. Results of overall survival (OS) prediction on TCGA-LUAD. Patient and slide counts in the Overall row denote totals across folds, whereas C-index values denote mean ± standard deviation across folds.

Fold	#Patients (#Slides)		C-Index ↑			
	Event	Censored	Bulk RNA	Img-Only	Img+PEKA	Img+COAST
0	33 (34)	59 (70)	0.5839	0.5019	0.5034	0.5146
1	33 (50)	58 (58)	0.6005	0.7134	0.6990	0.6986
2	32 (34)	59 (63)	0.5423	0.3949	0.4409	0.4978
3	32 (47)	59 (63)	0.6356	0.6635	0.6580	0.6560
4	32 (41)	59 (59)	0.6588	0.6488	0.6508	0.6473
Overall	162 (206)	294 (313)	0.6042 ± 0.05	0.5845 ± 0.13	0.5904 ± 0.11	0.6029 ± 0.09

combined with histology features, comparable to actual bulk RNA sequencing data. Future work will evaluate COAST on additional clinical cohorts and downstream tasks to better characterize fold-wise variability and further assess its robustness and generalizability. These findings suggest that context-aware differential learning is an effective strategy for enhancing histology-based spatial gene expression prediction.

Acknowledgments. This work was supported by a grant of the National Research Foundation of Korea (NRF) (No. RS-2025-00558322 and RS-2024-00397293).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abalo, X., Thrane, K., Ji, A., Mirzazadeh, R., Khavari, P.: Human squamous cell carcinoma, visium. Mendeley Data **10** (2021)
2. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nature medicine **30**(3), 850–862 (2024)
3. Chung, Y., Ha, J.H., Im, K.C., Lee, J.S.: Accurate spatial gene expression prediction by integrating multi-resolution features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11591–11600 (2024)
4. Cox, D.R.: Regression models and life-tables. Journal of the royal statistical society: Series B (methodological) **34**(2), 187–202 (1972)
5. Erickson, A., He, M., Berglund, E., Marklund, M., Mirzazadeh, R., Schultz, N., Kvastad, L., Andersson, A., Bergenstråhle, L., Bergenstråhle, J., et al.: Spatially resolved clonal copy number alterations in benign and malignant tissue. Nature **608**(7922), 360–367 (2022)
6. Ganguly, A., Chatterjee, D., Huang, W., Zhang, J., Yurovsky, A., Johnson, T.S., Chen, C.: 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. pp. 15611–15620 (2025)

7. He, B., Bergenstråhle, L., Stenbeck, L., Abid, A., Andersson, A., Borg, Å., Maaskola, J., Lundeberg, J., Zou, J.: Integrating spatial gene expression and breast tumour morphology via deep learning. Nature biomedical engineering **4**(8), 827–834 (2020)
8. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE international conference on computer vision. pp. 1501–1510 (2017)
9. Lee, J.C., Byeon, K., Song, B., Kim, K., Kwak, J.T.: Dior-vit: Differential ordinal learning vision transformer for cancer classification in pathology images. Medical Image Analysis **105**, 103708 (2025)
10. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature biomedical engineering **5**(6), 555–570 (2021)
11. Mejia, G., Ruiz, D., Cárdenas, P., Manrique, L., Vega, D., Arbeláez, P.: Enhancing gene expression prediction from histology images with spatial transcriptomics completion. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 91–101. Springer (2024)
12. Mirzazadeh, R., Andrusivova, Z., Larsson, L., Newton, P.T., Galicia, L.A., Abalo, X.M., Avijgan, M., Kvastad, L., Denadai-Souza, A., Stakenborg, N., et al.: Spatially resolved transcriptomic profiling of degraded and challenging fresh frozen samples. Nature Communications **14**(1), 509 (2023)
13. Oh Song, H., Xiang, Y., Jegelka, S., Savarese, S.: Deep metric learning via lifted structured feature embedding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4004–4012 (2016)
14. Pan, S., Chen, J., Secrier, M.: Teaching pathology foundation models to accurately predict gene expression with parameter efficient knowledge transfer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 605–613. Springer (2025)
15. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2337–2346 (2019)
16. Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: Film: Visual reasoning with a general conditioning layer. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
17. Qu, M., Wu, Y., Di, D., Gao, Y., Su, T., Song, Y., Fan, L.: Spatially gene expression prediction using dual-scale contrastive learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 574–584. Springer (2025)
18. Vicari, M., Mirzazadeh, R., Nilsson, A., Shariatgorji, R., Bjärterot, P., Larsson, L., Lee, H., Nilsson, M., Foyer, J., Ekvall, M., et al.: Spatial multimodal analysis of transcriptomes and metabolomes in tissues. Nature Biotechnology **42**(7), 1046–1050 (2024)
19. Villacampa, E.G., Larsson, L., Mirzazadeh, R., Kvastad, L., Andersson, A., Mollbrink, A., Kokaraki, G., Monteil, V., Schultz, N., Appelberg, K.S., et al.: Genome-wide spatial expression profiling in formalin-fixed tissues. Cell Genomics **1**(3) (2021)
20. Wang, H., Du, X., Liu, J., Ouyang, S., Chen, Y.W., Lin, L.: M2ost: Many-to-one regression for predicting spatial transcriptomics from digital pathology images. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 7709–7717 (2025)

21. Wang, X., Hua, Y., Kodirov, E., Hu, G., Garnier, R., Robertson, N.M.: Ranked list loss for deep metric learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5207–5216 (2019)
22. Xiao, X., Kong, Y., Li, R., Wang, Z., Lu, H.: Transformer with convolution and graph-node co-embedding: An accurate and interpretable vision backbone for predicting gene expressions from local histopathological image. Medical Image Analysis **91**, 103040 (2024)
23. Yang, Y., Hossain, M.Z., Stone, E., Rahman, S.: Spatial transcriptomics analysis of gene expression prediction using exemplar guided graph neural network. Pattern Recognition **145**, 109966 (2024)
24. Zeng, Y., Wei, Z., Yu, W., Yin, R., Yuan, Y., Li, B., Tang, Z., Lu, Y., Yang, Y.: Spatial transcriptomics prediction from histology jointly through transformer and graph neural networks. Briefings in Bioinformatics **23**(5), bbac297 (2022)