Spatially Gene Expression Prediction using Dual-Scale Contrastive Learning

Mingcheng Qu[1], Yuncong Wu[2], Donglin Di[3], Yue Gao[3], Tonghua Su[1], Yang Song[4], and Lei Fan[4]✉

[1] Faculty of Computing, Harbin Institute of Technology, Harbin, China
[2] School of Astronautics, Harbin Institute of Technology, Harbin, China
[3] School of Software, Tsinghua University, Beijing, China
[4] School of Computer Science and Engineering, UNSW, Sydney, Australia
lei.fan1@unsw.edu.au

Abstract. Spatial transcriptomics (ST) provides crucial insights into tissue micro-environments, but is limited to its high cost and complexity. As an alternative, predicting gene expression from pathology whole slide images (WSI) is gaining increasing attention. However, existing methods typically rely on single patches or a single pathology modality, neglecting the complex spatial and molecular interactions between target and neighboring information (*e.g.*, gene co-expression). This leads to a failure in establishing connections among adjacent regions and capturing intricate cross-modal relationships. To address these issues, we propose NH²2ST, a framework that integrates spatial context and both pathology and gene modalities for gene expression prediction. Our model comprises a query branch and a neighbor branch to process paired target patch and gene data and their neighboring regions, where cross-attention and contrastive learning are employed to capture intrinsic associations and ensure alignments between pathology and gene expression. Extensive experiments on six datasets demonstrate that our model consistently outperforms existing methods, achieving over 20% in PCC metrics. Codes are available at https://github.com/MCPathology/NH2ST.

Keywords: Gene expression · Spatial transcriptomics · Contrastive learning · Cross-attention.

1 Introduction

Gene expression plays a crucial role in understanding cellular functions and disease mechanisms. Recently, spatial transcriptomics (ST) [15] has emerged to integrate both spatial coordinates and transcriptomic signatures [4], enabling spatially resolved gene expression profiling [27]. Despite its advantages, the high cost and technical complexity limit its scalability [24]. This highlights the need for computational approaches capable of predicting spatially resolved gene expression directly from whole slide images (WSIs) [14,6,5].

✉ Corresponding author

Earlier studies [9,26] treated individual patches as input with gene expression as labels. Some approaches then incorporated spatial context, such as HisToGene [18] and mclSTExp [16]. TRIPLEX [3] integrated features from the entire neighborhood patch. Hist2ST [36] further improved upon this by using graphs to capture neighborhood relationships. However, these methods primarily focus on capturing features from a broader context coarsely (*e.g.*, extracting features from larger pathology patches) without explicitly modeling inter-patch relationships, limiting their ability to fully capture the complex spatial and molecular interactions present in pathology images [36,20]. Since ST data inherently displays strong spatial continuity [23], and gene expression in adjacent locations is often correlated due to biological factors like cell-cell interactions [1] and microenvironment [18,28]. Consequently, leveraging the spatial context from neighboring patches and establishing their interactions are essential.

On the other hand, most previous methods [9,18,36] rely on only pathology inputs for predicting the gene expression. Some studies [33] explored example-based learning to retrieve relevant samples, while recent works [32,16] employed contrastive learning [22,7] to align these pathology and gene expression modalities. However, while contrastive learning enforces alignment between pathology image features and gene expression features at the same spatial location, it does not fully consider the interactions between the two modalities. Pathology features are often associated with specific gene expressions (*e.g.*, invasive margins strongly correlate with PD-L1 expression [29]), requiring a more dynamic and integrative approach to aggregate information from both modalities [21].

To address these challenges, we aim to integrate spatial neighborhood information and multimodal interactions between pathology and gene expression, where hypergraph learning [8,13] is employed to capture neighboring relationships, while cross-attention [31] and contrastive learning [22,13] are used to model interactions and ensure alignment between the two modalities. In this paper, we propose NH^{2}2ST, Neighbor-guided Hypergraph Histopathology to ST gene expression. It consists of two branches: 1) A query branch processes the paired target patch and ST data, where cross-attention and cross-modal contrastive learning are applied to refine and align pathology-gene features. 2) A neighbor branch extends this by incorporating the target patch along with its neighboring regions and paired ST data, leveraging hypergraphs to model interactions among different patches and their ST spots. The neighbor branch plays an auxiliary role, supporting the query branch in training more robust feature extractors. Meanwhile, the pathology extractor in the query branch is further enhanced with a prediction translator to generate gene expression values during inference.

Our contributions can be summarized as: We propose a dual-scale framework for predicting spatially resolved gene expression from pathology patches, employing cross-attention and contrastive learning to enhance alignment between the pathology and gene modalities. We leverage hypergraph learning to model the interactions among different patches for both modalities, capturing complex spatial and molecular interactions effectively. Extensive experiments on six publicly available datasets demonstrate that our model outperforms advanced methods.

2 Method

2.1 Overview

Preliminaries. Given a WSI X represented as M non-overlapping tissue patches, $\{x_i\}_{i=1}^{M}$, where each patch $x_i \in \mathbb{R}^{h\times w\times 3}$ corresponds to a ST spot detected by a probe, paired with a gene expression vector $y_i \in \mathbb{R}^n$ capturing the normalized expression levels of n genes. To mitigate technical noise and enhance signal quality, we followed previous studies [9,3] by selecting n=250 most highly expressed genes, applying a logarithmic transformation to their expression values. The objective is to learn a mapping function $\phi : x_i \rightarrow y_i$ to predict gene expression from corresponding patches.

Framework. NH^{2}2ST consists of a query branch and a neighbor branch (see Fig. 1). The query branch extracts patch features $h_s^p \in \mathbb{R}^N$ and gene features $h_s^g \in \mathbb{R}^N$ from pathology-gene pairs $\langle x_i, y_i\rangle$ using pathology and ST encoders. These features are mutually refined via cross-attention and aligned through contrastive learning. In the neighbor branch, the patch x_i is considered alongside its K nearest neighboring patches and corresponding gene data. Two hypergraphs are constructed based on feature similarities and spatial proximity, and these graph representations are further refined and aligned through cross-attention and contrastive learning. Importantly, the patch features h_s^p in the query branch are further processed by a translator module to generate the final gene expression prediction $\hat{y}_i$. Both branches share the pathology and ST encoders, with the neighbor branch serving as an auxiliary component. It enhances the pathology encoder while capturing relationships between multi-scale pathology and gene expression.

Encoders. Following previous studies [3,32,34], the pathology encoder $\phi_p : \mathbb{R}^{h\times w\times 3} \rightarrow \mathbb{R}^N$ is a truncated pre-trained ResNet18 [10], extracting a 224×224 patch x_i to generate an N-dimensional feature $h^p \in \mathbb{R}^N$. The ST encoder $\phi_g : \mathbb{R}^n \rightarrow \mathbb{R}^N$ consists of two fully connected layers, widely used to capture non-linear interactions, producing gene features $h^g \in \mathbb{R}^N$.

2.2 Query Branch

This branch learns to predict gene expression values $\hat{y}$ from the patch x. It comprises a cross-attention module, cross-modal contrastive learning, and a prediction translator. Specifically, given the extracted feature pair $\langle h_s^p, h_s^g\rangle$, the refined pathology features z_s^p are computed as follows:

$$z_s^p = \phi_{ca}^s(h_s^p, h_s^g) = \text{softmax}\left(\frac{(\mathbf{W}_q^s h_s^p)(\mathbf{W}_k^s h_s^g)^T}{\sqrt{d_k}}\right)(\mathbf{W}_v^s h_s^g), \tag{1}$$

where the cross-attention function $\phi_{ca}^s(*, \circ)$ refines $*$ using guidance from $\circ$, $\mathbf{W}_q^s$, $\mathbf{W}_k^s$ and $\mathbf{W}_v^s$ are learnable projection matrices, d_k represents the dimensionality of the key vector projection $\mathbf{W}_k^s h_s^g$. Similarly, the refined gene features z_s^g are obtained as $z_s^g = \phi_{ca}^s(h_s^g, h_s^p)$. This bi-directional mechanism [31] enables each

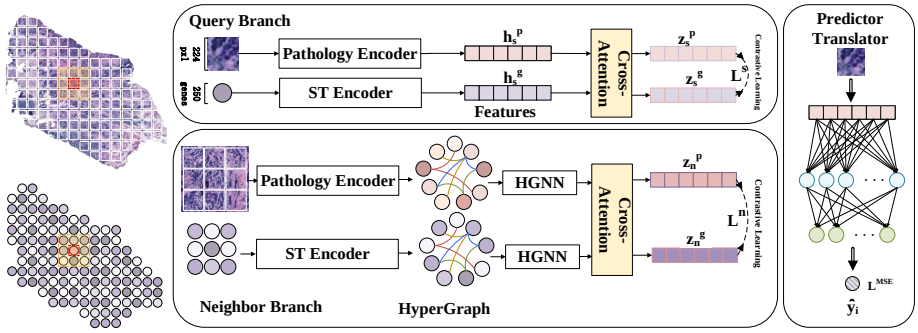

Fig. 1. The architecture of NH²2ST. Our model consists of two key components: a query branch and a neighbor branch. It employs dual-level contrastive losses to strengthen the intrinsic connections between the two modalities, facilitating the final gene expression prediction.

modality (*e.g.*, pathology or gene) to attend to relevant patterns in the other, establishing semantic correspondences and aligning histological structures with transcriptomic profiles.

The refined feature pair $\langle z_s^p, z_s^g \rangle$ is further aligned explicitly using cross-modal contrastive learning, where pairs from the same spatial location are treated as positive, while unpaired features serve as negative. The contrastive objective $\mathcal{L}^s$ is defined using the InfoNCE loss [17]:

$$\mathcal{L}^s = \mathcal{L}(\{z_s^p\}^B, \{z_s^g\}^B) = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\mathcal{P}^+(z_{s\,i}^p, z_{s\,i}^g)}{\mathcal{P}^+(z_{s\,i}^p, z_{s\,i}^g) + \sum_{j=1}^{B-1} \mathcal{P}^-(z_{s\,i}^p, z_{s\,j}^g)}, \quad (2)$$

where B is the batch size, $\mathcal{P}^+(\cdot,\cdot) = \exp(\mathrm{sim}(\cdot,\cdot)/\tau)$ denotes the similarity of positive pairs, $\mathcal{P}^-(\cdot,\cdot) = \exp(\mathrm{sim}(\cdot,\cdot)/\tau)$ represents the similarity of negative pairs, and τ is the temperature coefficient. Contrastive learning aligns pathology and gene features in a shared latent space while preserving biological relationships. It enhances discriminative representation by clustering same-spatial features while separating irrelevant ones, improving robustness to biological and technical noise (*e.g.*, cell-type heterogeneity [19] and batch effects [30]).

Both cross-attention and contrastive learning are used to capture the intrinsic relationship between pathology and gene expression. Our goal is to predict gene expression using only pathology patches during inference. To achieve this, we introduce a prediction translator $\phi_t : \mathbb{R}^N \to \mathbb{R}^n$ that converts the extracted pathology features h_s^p into predicted gene expression values $\hat{y} = \phi_t(h_s^p)$. The translator consists of two fully connected layers (opposite to the ST encoder), and is optimized using the Mean Squared Error (MSE) loss [3,33], denoted as: $\mathcal{L}_{MSE} = \frac{1}{B}\sum_{i=1}^{B}(y_i - \hat{y}_i)^2$.

In the training, the pathology encoder ϕ_p is simultaneously driven by the translator and contrastive learning. This creates a "loop" [35]. Rather than causing instability, it typically enhances multimodal learning by enforcing representa-

tions that are both predictive (for the translator) and coherent across modalities (for contrastive alignment).

2.3 Neighbor Branch

This branch serves as an auxiliary component to assist the query branch in training pathology encoders ϕ_p by effectively utilizing neighbor information. It consists of hypergraph representation, a cross-attention module, and cross-modal contrastive learning. Given a patch x, we select its K nearest spatial neighboring patches $\{x_j\}_{j=1}^{K}$ and corresponding ST data. We then use a hypergraph to extract features from these neighboring patches. Unlike traditional graphs [36], hypergraphs connect multiple nodes within a single edge, capturing higher-order relationships. This is inspired by groups of co-expressed genes extending beyond pairwise edges [25], leading to a richer structure and more robust alignment.

Take pathology patches as an example, we construct a graph $\mathcal{G}^p = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{v_1, \ldots, v_K\}$ represents the node set and $\mathcal{E} = \{e_1, \ldots, e_K\}$ denotes the hyperedge set. Each node v_k corresponds to a patch x_k. Each hyperedge e_j has a degree of τ, connecting node v_j with its $\tau - 1$ most similar nodes. The similarity between nodes is calculated as follows:

$$e_j = \{v_j\} \cup \left\{ v_k \;\middle|\; k \in \arg\max_{k \neq j}^{\tau-1} \mathrm{sim}_{jk}(h_{nj}^p, h_{nk}^p) \right\}, \tag{3}$$

where $\mathrm{sim}(\cdot, \cdot)$ represents the cosine similarity. Similarly, a gene hypergraph $\mathcal{G}^g$ is constructed. Both hypergraphs are processed through separate L-layer HGNNs and a pooling layer, yielding paired graph representations $\langle h_n^p \in \mathbb{R}^N, h_n^g \in \mathbb{R}^N \rangle$. The hypergraph models both the spatial morphology of pathology patches and gene co-expression patterns [25], capturing relationships between a patch and its surrounding tissue structures, as well as between gene expression profiles in spatially adjacent regions.

Then, similar to the query branch, the cross-attention module ϕ_{ca}^n is applied to obtain refined graph features $\langle z_n^p, z_n^g \rangle$ as followed by:

$$z_n^p = \phi_{ca}^n(h_n^p, h_n^g), z_n^g = \phi_{ca}^n(h_n^g, h_n^p). \tag{4}$$

By incorporating cross-attention at the neighbor level through hypergraph, we enrich contextual integration while capturing structural intricacies. This enhances the alignment of multimodal relationships. The refined features $\langle z_n^p, z_n^g \rangle$ are then used for contrastive learning as: $\mathcal{L}^n = \mathcal{L}(\{z_n^p\}^B, \{z_n^g\}^B)$. By aligning the two modalities at the neighborhood level, it enhances the capture of broader spatial and co-expression interactions.

2.4 Training and Inference

The training objective consists of contrastive learning losses from both the query and neighbor branches, along with the translation loss in the query branch. The total loss $\mathcal{L}$ is computed as:

$$\mathcal{L} = \lambda_1 \cdot \mathcal{L}^s + \lambda_2 \cdot \mathcal{L}^n + \mathcal{L}_{MSE}, \tag{5}$$

Table 1. Comparison of our model with advanced models on six datasets.

Models	MSE (↓)	MAE (↓)	PCC (↑)	MSE (↓)	MAE (↓)	PCC (↑)
	STNet dataset			Skin		
ST-Net [9]	0.209±0.02	0.349±0.02	0.223±0.10	0.294±0.07	0.428±0.05	0.382±0.08
EGN [33]	0.192±0.02	0.337±0.04	0.203±0.09	0.281±0.08	0.418±0.06	0.388±0.06
BLEEP [32]	0.235±0.02	0.369±0.02	0.193±0.10	0.297±0.08	0.430±0.04	0.396±0.08
HistoGene [18]	0.199±0.03	0.335±0.04	0.201±0.09	0.294±0.07	0.415±0.07	0.406±0.10
His2ST [36]	0.181±0.07	0.333±0.02	0.199±0.07	0.291±0.16	0.924±0.29	0.353±0.07
TRIPLEX [3]	0.202±0.07	0.343±0.02	0.352±0.10	0.268±0.09	0.404±0.07	**0.490±0.07**
NH²2ST (our)	**0.161±0.03**	**0.311±0.02**	**0.572±0.10**	**0.208±0.15**	**0.349±0.08**	0.478±0.11
Models	INT(HEST-1k)			ZEN(HEST-1k)		
ST-Net [9]	**0.138±0.02**	**0.278±0.04**	0.213±0.06	**0.079±0.01**	**0.159±0.01**	0.155 ± 0.01
BLEEP [32]	0.403±0.04	0.333±0.06	0.104±0.07	0.172±0.01	0.200±0.01	0.229±0.20
TRIPLEX [3]	0.221±0.01	0.390±0.03	0.184±0.01	0.099±0.02	0.150±0.01	**0.242±0.03**
NH²2ST (our)	0.220±0.03	0.388±0.03	**0.368±0.02**	0.108±0.01	0.193±0.01	0.216±0.02
Models	PCW(STimage-1K4M)			Mouse(STimage-1K4M)		
ST-Net [9]	0.034±0.01	0.139±0.01	0.466±0.02	0.041±0.01	0.147±0.01	0.256±0.01
BLEEP [32]	0.166±0.09	0.295±0.05	0.241±0.04	0.085±0.01	0.219±0.01	0.165±0.01
TRIPLEX [3]	0.039±0.01	0.144±0.01	0.512±0.01	0.152±0.03	0.304±0.01	0.200±0.01
NH²2ST (our)	**0.029±0.01**	**0.129±0.02**	**0.560±0.03**	**0.039±0.02**	**0.107±0.01**	**0.593±0.01**

where λ_1 and λ_2 are balancing weights. During inference, since gene expression data is unavailable, only the Predictor Translator is used to predict gene expression values $\hat{y}_i = \phi_g(\phi_p(x_i))$ based on input patch x_i.

3 Experiments

We conducted experiments on: ST-Net [9], Skin [12], two datasets from STimage-1K4M [2], and two datasets from HEST-1k [11]. We followed the previous data split settings [3,11]. Image patches were resized to 224 × 224 pixels, and 250 highly expressed genes were selected. Adam optimizer was employed with an initial learning rate of 0.0001, dynamically adjusted with a step size of 50 and a decay rate of 0.9. The batch size was 8 and the training epoch was 20.

We evaluated model performance using three metrics [3]: Mean Squared Error (MSE) for error magnitude, Mean Absolute Error (MAE) for average absolute differences, and Pearson Correlation Coefficient (PCC) for linear correlation. Lower MSE and MAE indicate higher accuracy, while higher PCC shows stronger consistency between predictions and actual values. We conducted k-fold cross-validation and reported results as mean ± standard deviation (STD).

3.1 Results and Discussion

We compared our model against several advanced methods (see Table 1). Results for the ST-Net and Skin datasets were quoted from TRIPLEX [3], and results for

Table 2. Left: Ablation studies of branch construction. Variables tested include: query ($Q.$) and neighbor ($N.$) branch, cross-attention ($CA.$), contrastive learning ($CL.$), graph types including graph (G.) and hypergraph($HG.$), self-attention ($SA.$). **Right:** Results for different K nearest patches and L-layer HGNN in hypergraph.

Branch	$CA.$	$CL.$	MSE (↓)	MAE (↓)	PCC (↑)
$Q.$			0.147±0.06	0.331±0.06	0.320±0.22
$Q.$	$SA.$		0.087±0.01	0.224±0.02	0.155±0.01
$Q.$		✓	0.781±0.06	0.809±0.11	0.424±0.03
$Q.$	✓	✓	0.093±0.04	0.229±0.06	0.329±0.06
$Q. + N.$	✓		0.106±0.07	0.195±0.05	0.468±0.05
$Q. + N.$		✓	0.597±0.05	0.703±0.04	0.451±0.03
$Q. + N.$ ($G.$)	✓	✓	0.043±0.02	0.154±0.06	0.411±0.02
$Q. + N.$ ($HG.$)	$SA.$	✓	0.035±0.01	0.143±0.01	0.526±0.03
NH²2ST			**0.029±0.01**	**0.129±0.02**	**0.560±0.03**

Configuration	MSE (↓)	MAE (↓)	PCC (↑)
K=4	0.034±0.01	0.142±0.01	0.525±0.03
K=8	**0.033±0.01**	**0.141±0.01**	**0.530±0.03**
K=16	0.035±0.01	0.144±0.01	0.529±0.02
K=25	0.036±0.01	0.146±0.01	0.526±0.03
L=1	0.062±0.01	0.191±0.02	0.469±0.04
L=2	**0.033±0.01**	**0.141±0.01**	**0.530±0.02**
L=3	0.093±0.04	0.231±0.05	0.256±0.14
L=4	0.130±0.05	0.276±0.06	0.207±0.09
NH²2ST	0.029±0.01	0.129±0.02	0.560±0.03

other datasets were reproduced using the released codes. Our model achieved the best performance across three metrics on STNet, Skin, and two STimage-1K4M datasets, and comparable results on HEST-1k. Notably, PCC results showed significant improvements over previous methods across all datasets. We attribute these gains to the neighbor branch with graph learning, which explicitly models relationships between the query image and adjacent regions, and the use of contrastive learning and cross-attention, which enhances interactions between pathology and gene modalities.

3.2 Ablation Studies

Branch Construction. We evaluated the impact of different branch components on PCW dataset (see Table 2). When using only the query branch, incorporating self-attention ($SA.$) improved MSE and MAE but reduced PCC by 0.165, while using contrastive learning ($CL.$) increased PCC by 0.104 but significantly raised errors. Integrating $CL.$ with cross-attention $CA.$ improved all three metrics. For the neighbor branch, a similar trend was observed, using only $CA. + CL.$ led to significant performance gains. Additionally, we examined the effect of graph ($G.$) or hypergraphs ($HG.$) with self-attention in the neighbor branch. While both performed reasonably well, they were inferior to our proposed method. These results highlight that the combination of contrastive learning and cross-attention is crucial, as cross-attention facilitates interactions between modalities.

Moreover, we tested different neighbor sizes and HGNN layers, observing the best performance at $K = 8$ and $L = 2$. We attribute this to the fact that excessive neighbors introduce too much regional information, while more HGNN layers lead to over-smoothing of node information.

Contrastive Learning. We explored the impact of different contrastive learning on the PCW dataset (see Table 3). We observed that all three metrics achieved optimal results with a batch size of 8, particularly PCC, which reached 0.550. As the batch size increased further, no significant improvements were observed. Regarding the balance parameters, setting $\lambda_1 = 1$ and $\lambda_2 = 0.5$ yielded

Table 3. Ablation study of contrastive learning. Investigating the impact of batch size (B), loss balancing hyperparameters (λ_1, λ_2), and temperature coefficient τ.

B	MSE (↓)	MAE (↓)	PCC (↑)	(λ1,λ2)	MSE (↓)	MAE (↓)	PCC (↑)	τ	MSE (↓)	MAE (↓)	PCC (↑)
4	0.182±0.03	0.300±0.02	0.397±0.01	(0,1)	**0.042±0.01**	**0.150±0.01**	0.470±0.04	0.025	0.191±0.03	0.333±0.02	0.228±0.19
8	**0.030±0.01**	**0.132±0.01**	**0.550±0.02**	(1,0)	0.093±0.04	0.229±0.05	0.353±0.08	0.05	**0.039±0.01**	**0.152±0.01**	**0.505±0.04**
16	0.034±0.01	0.144±0.01	0.534±0.02	(0.5,1)	0.050±0.01	0.173±0.010	0.494±0.02	0.1	0.062±0.01	0.191±0.02	0.469±0.04
32	0.062±0.01	0.191±0.02	0.469±0.03	(1,0.5)	0.043±0.01	0.160±0.01	**0.508±0.03**	0.15	0.043±0.01	0.158±0.01	0.483±0.03
64	0.068±0.01	0.284±0.09	0.482±0.06	(1,1)	0.061±0.01	0.193±0.02	0.469±0.04	0.2	0.039±0.01	0.153±0.01	0.491±0.02
NH²2ST	0.029±0.01	0.129±0.02	0.560±0.03	NH²2ST	0.029±0.01	0.129±0.02	0.560±0.03	NH²2ST	0.029±0.01	0.129±0.02	0.560±0.03

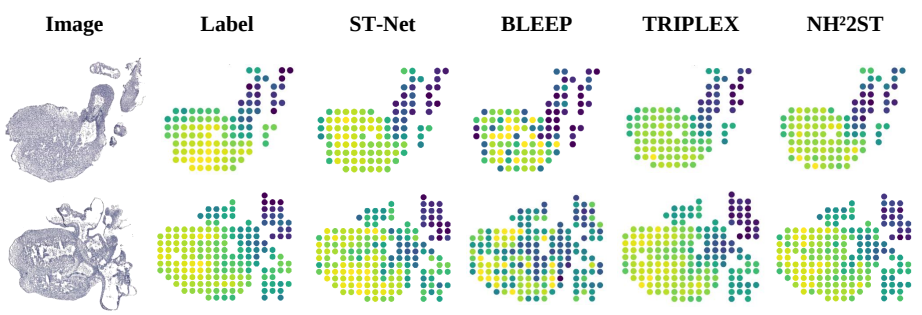

Fig. 2. Visualization of expression values on the PCW Dataset. From left to right: raw WSI, labels, ST-Net, BLEEP, TRIPLEX and our NH²2ST.

the best results. While $\lambda_1 = 0$ and $\lambda_2 = 1$ slightly improved MSE and MAE, it led to 7.4% lower PCC compared to our chosen configuration. This indicates a crucial trade-off between the two loss functions in determining the final performance. For the temperature coefficient, a value of 0.05 yielded the best performance. Higher values resulted in slightly reduced performance but remained relatively stable.

Visualization. We conducted a visualization comparison of our model against other methods on the PCW dataset (see Fig. 2). Compared to BLEEP (contrastive learning-based) and TRIPLEX (neighbor-based), our model exhibited higher visual consistency with the true values, particularly in highly expressed regions, demonstrating its effectiveness in accurate prediction.

4 Conclusion

We present NH²2ST that integrates dual-scale spatial modeling and contrastive learning to predict spatially resolved gene expression from whole slide images. It leverages hypergraphs to capture contextual relationships among adjacent regions and employs cross-attention contrastive learning to refine and align pathology and gene expression modalities. Experiments on six publicly available datasets demonstrate the superior performance of our framework.

Limitation. NH²2ST utilizes only the query branch during inference, as experiments show that incorporating the neighborhood branch may introduce complexity or noise, leading to performance degradation. Additionally, the frame-

work relies on well-aligned pathology images and ST data, with misalignment or noise could impact accuracy. Future work may explore effective integration of neighborhood information during inference and develop robust preprocessing techniques to mitigate data misalignment and noise.

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

References

1. Armingol, E., Officer, A., Harismendy, O., et al.: Deciphering cell–cell interactions and communication from gene expression. Nature Reviews Genetics **22**(2), 71–88 (2021)
2. Chen, J., Zhou, M., Wu, W., et al.: Stimage-1k4m: A histopathology image-gene expression dataset for spatial transcriptomics. arXiv preprint arXiv:2406.06393 (2024)
3. Chung, Y., Ha, J.H., Im, K.C., et al.: Accurate spatial gene expression prediction by integrating multi-resolution features. In: CVPR. pp. 11591–11600 (2024)
4. Eng, C.H.L., Lawson, M., Zhu, Q., et al.: Transcriptome-scale super-resolved imaging in tissues by rna seqfish+. Nature **568**(7751), 235–239 (2019)
5. Fan, L., Sowmya, A., Meijering, E., Song, Y.: Learning visual features by colorization for slide-consistent survival prediction from whole slide images. In: MICCAI. pp. 592–601. Springer (2021)
6. Fan, L., Sowmya, A., Meijering, E., et al.: Cancer survival prediction from whole slide images with self-supervised learning and slide consistency. IEEE Transactions on Medical Imaging **42**(5), 1401–1412 (2022)
7. Fan, L., Sowmya, A., Meijering, E., et al.: Fast ff-to-ffpe whole slide image translation via laplacian pyramid and contrastive learning. In: MICCAI. pp. 409–419. Springer (2022)
8. Gao, Y., Zhang, Z., Lin, H., et al.: Hypergraph learning: Methods and practices. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(5), 2548–2566 (2020)
9. He, B., Bergenstråhle, L., Stenbeck, L., et al.: Integrating spatial gene expression and breast tumour morphology via deep learning. Nature biomedical engineering **4**(8), 827–834 (2020)
10. He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
11. Jaume, G., Doucet, P., Song, A., et al.: Hest-1k: A dataset for spatial transcriptomics and histology image analysis. NeurIPS **37**, 53798–53833 (2025)
12. Ji, A.L., Rubin, A.J., Thrane, K., et al.: Multimodal analysis of composition and spatial architecture in human squamous cell carcinoma. Cell **182**(2), 497–514 (2020)
13. Jing, W., Wang, J., Di, D., et al.: Multi-modal hypergraph contrastive learning for medical image segmentation. Pattern Recognition **165**, 111544 (2025)
14. Lee, M.: Recent advancements in deep learning using whole slide imaging for cancer prognosis. Bioengineering **10**(8), 897 (2023)
15. Maynard, K., Jaffe, A., Martinowich, K.: Spatial transcriptomics: putting genome-wide expression on the map. Neuropsychopharmacology **45**(1), 232 (2019)

16. Min, W., Shi, Z., Zhang, J., et al.: Multimodal contrastive learning for spatial gene expression prediction using histology images. Briefings in Bioinformatics **25**(6), bbae551 (2024)
17. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
18. Pang, M., Su, K., Li, M.: Leveraging information in spatial transcriptomics to predict super-resolution gene expression from histology images in tumors. BioRxiv pp. 2021–11 (2021)
19. Qi, L., Teschendorff, A.E.: Cell-type heterogeneity: Why we should adjust for it in epigenome and biomarker studies. Clinical Epigenetics **14**(1), 31 (2022)
20. Qu, M., Wu, Y., Di, D., et al.: Boundary-guided learning for gene expression prediction in spatial transcriptomics. In: BIBM. pp. 445–450. IEEE (2024)
21. Qu, M., Yang, G., Di, D., et al.: Multimodal cancer survival analysis via hypergraph learning with cross-modality rebalance. IJCAI (2025)
22. Radford, A., Kim, J.W., Hallacy, C., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763. PMLR (2021)
23. Rao, A., Barkley, D., França, G.S., et al.: Exploring tissue architecture using spatial transcriptomics. Nature **596**(7871), 211–220 (2021)
24. Shen, X., Zhao, Y., Wang, Z., et al.: Recent advances in high-throughput single-cell transcriptomics and spatial transcriptomics. Lab on a Chip **22**(24), 4774–4791 (2022)
25. Song, W.M., Zhang, B.: Multiscale embedded gene co-expression network analysis. PLoS computational biology **11**(11), e1004574 (2015)
26. Srinivas, A.A., Jaroensri, R., Wulczyn, E., et al.: Estrogen receptor gene expression prediction from h&e whole slide images. medRxiv pp. 2024–04 (2024)
27. Ståhl, P.L., Salmén, F., Vickovic, S., et al.: Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science **353**(6294), 78–82 (2016)
28. Tang, Q., Fan, L., Pagnucco, M., Song, Y.: Prototype-based image prompting for weakly supervised histopathological image segmentation. CVPR (2025)
29. Topalian, S.L., Taube, J.M., Anders, R.A., et al.: Mechanism-driven biomarkers to guide immune checkpoint blockade in cancer therapy. Nature Reviews Cancer **16**(5), 275–287 (2016)
30. Vaessen, N., van Leeuwen, D.A.: The effect of batch size on contrastive self-supervised speech representation learning. arXiv preprint arXiv:2402.13723 (2024)
31. Wei, X., Zhang, T., Li, Y., et al.: Multi-modality cross attention network for image and sentence matching. In: CVPR. pp. 10941–10950 (2020)
32. Xie, R., Pang, K., Chung, S., et al.: Spatially resolved gene expression prediction from histology images via bi-modal contrastive learning. NeurIPS **36** (2024)
33. Yang, Y., Hossain, M.Z., Stone, E.A., et al.: Exemplar guided deep neural network for spatial transcriptomics analysis of gene expression prediction. In: WACV. pp. 5039–5048 (2023)
34. Yi, K., Chen, J., Wang, Y.G., et al.: Approximate equivariance so (3) needlet convolution. In: Topological, Algebraic and Geometric Learning Workshops 2022. pp. 189–198. PMLR (2022)
35. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: Coca: Contrastive captioners are image-text foundation models. arXiv preprint arXiv:2205.01917 (2022)
36. Zeng, Y., Wei, Z., Yu, W., et al.: Spatial transcriptomics prediction from histology jointly through transformer and graph neural networks. Briefings in Bioinformatics **23**(5), bbac297 (2022)