MMAP: A Multi-Magnification and Prototype-Aware Architecture for Predicting Spatial Gene Expression

Hai Dang Nguyen[1 [0009−0008−5752−1049]⋆], Nguyen Dang Huy Pham[2⋆⋆], The Minh Duc Nguyen[1], Dac Thai Nguyen[1], Hang Thi Nguyen[3⋆⋆⋆], and Duong M. Nguyen[4†]

[1] Institute for AI Innovation and Societal Impact, Hanoi University of Science and Technology, Hanoi, Vietnam
[2] Amsterdam High School for the Gifted, Hanoi, Vietnam
[3] Anatomic Pathology Division, Laboratory Department, Vinmec Times City International Hospital, Vinmec Healthcare System, Hanoi, Vietnam
[4] University of Illinois Urbana-Champaign, Illinois, USA
nmduongg@illinois.edu

Abstract. Spatial Transcriptomics (ST) enables the measurement of gene expression while preserving spatial information, offering critical insights into tissue architecture and disease pathology. Recent developments have explored the use of hematoxylin and eosin (H&E)-stained whole-slide images (WSIs) to predict transcriptome-wide gene expression profiles through deep neural networks. This task is commonly framed as a regression problem, where each input corresponds to a localized image patch extracted from the WSI. However, predicting spatial gene expression from histological images remains a challenging problem due to the significant modality gap between visual features and molecular signals. Recent studies have attempted to incorporate both local and global information into predictive models. Nevertheless, existing methods still suffer from two key limitations: (1) insufficient granularity in local feature extraction, and (2) inadequate coverage of global spatial context. In this work, we propose a novel framework, MMAP (Multi-MAgnification and Prototype-enhanced architecture), that addresses both challenges simultaneously. To enhance local feature granularity, MMAP leverages multi-magnification patch representations that capture fine-grained histological details. To improve global contextual understanding, it learns a set of latent prototype embeddings that serve as compact representations of slide-level information. Extensive experimental results demonstrate that MMAP consistently outperforms all existing state-of-the-art methods across multiple evaluation metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), and Pearson Correlation Coefficient (PCC).

⋆ Co-first author.
⋆⋆ Co-first author.
⋆⋆⋆ Co-corresponding author.
† Co-corresponding author.

Keywords: Spatial transcriptomics · Histology · Deep learning · Prototype · Multi-magnification.

1 Introduction

Background. Spatial Transcriptomics (ST) quantifies mRNA expression for a defined set of genes across a tissue sample by segmenting it into discrete "spots". Unlike bulk RNA sequencing, which measures gene expression across an entire tissue and misses intra-sample heterogeneity and spatial relationships, or single-cell RNA sequencing (scRNA-seq), which captures cell-level heterogeneity but loses spatial context due to cell isolation, ST methods preserve both molecular and spatial information. Advanced ST techniques [1, 3, 12, 13, 19, 30] enable analysis at varying resolutions, from multi-cell tissue segments [13] to single cells [19] or subcellular regions. Despite their scientific potential, ST methods remain costly, requiring specialized expertise, equipment, and reagents. To this end, deep neural networks have emerged as a promising solution to inferring gene expression profiles directly from histological images.

Deep learning-based spatial transcriptomics prediction. Early efforts in gene expression prediction commonly framed the task as a regression problem, wherein models were trained to estimate gene expression levels from individual image patches. These approaches typically employed convolutional neural networks (CNNs) or transformer-based architectures to learn visual representations from histopathological inputs. In this context, each tissue spot was encoded using deep features extracted from intermediate layers of pretrained models, such as ResNet18 or ResNet101 [6, 26], which effectively capture both fine-grained morphological cues and higher-order structural patterns. Building upon this foundation, subsequent methods have sought to enhance predictive performance by leveraging these image-derived feature representations for spatial transcriptomics (ST) data. For example, ST-Net [8] applies a transfer learning strategy, fine-tuning a DenseNet121 model pretrained on ImageNet to predict gene expression from histology images. THItoGene [16] introduces a more sophisticated architecture based on dynamic convolutional layers and capsule networks to infer RNA-Seq profiles from whole-slide images (WSIs). Extending this line of work, HisToGene [17] employs a Vision Transformer (ViT) to model patch-level correlations across WSIs, enabling gene expression prediction informed by global spatial context. Despite their promise, these models are fundamentally limited by their narrow focus on isolated patches, ignoring the broader spatial dependencies inherent in whole-slide images. Recent studies [9, 25, 29] integrate pathology images and spatial information, employing graph neural networks (GNNs) to model complex spot interactions and spatial relationships within tissues.

Limitations of existing approaches and our solution. However, existing methods still fall short of modeling long-range dependencies and often overlook the hierarchical nature of histopathological information. In practice, gene expression in a given region is influenced not only by its immediate surroundings but also by distant tissue context, sometimes requiring analysis at the whole-slide

level to detect disease-specific patterns. Meanwhile, current models typically extract local features only at the patch level, missing fine-grained morphological signals that often become apparent only under higher magnifications (e.g. 20x, 40×). To tackle these challenges, we propose a novel framework that targets both the granularity and spatial scope of information extraction. First, to enhance fine-grained local representation, we introduce a multi-magnification strategy using random cropping to generate sub-patches, allowing the model to access high-resolution details. Second, to capture broader spatial context while maintaining scalability, we learn a set of representative prototype embeddings that summarize global tissue-level patterns. These prototypes act as contextual anchors, enabling the model to integrate both local and global information for more accurate gene expression prediction. To realize these ideas, we introduce MMAP, a novel deep neural network that integrates **M**ulti-**MA**gnification features and builds a **P**rototype bank to seamlessly combine local and global information for gene expression prediction, while maintaining computational efficiency. MMAP operates in two key stages. Firstly, it generates spot-level features by leveraging multi-magnification views to capture detailed local context. Next, these features are used to create a prototype bank - a compact set of representative embeddings that encapsulate the entire WSI. This bank serves as a condensed representation of the WSI, providing rich global context without the need to process thousands of image patches. By employing a cross-attention mechanism, MMAP effectively combines multi-magnification features and models interactions between spots within the tissue. This approach enhances prediction accuracy by utilizing both magnification-specific details and spatial relationships, all while minimizing computational overhead.

Our contributions. The key contributions are summarized as follows:

- We present a novel framework for predicting spatial gene expression levels directly from WSIs. Our approach is designed to capture local histological features at multiple levels of granularity while simultaneously leveraging global contextual information to improve spatial relationship modeling. Crucially, the proposed method maintains computational efficiency and scalability, making it well-suited for the analysis of high-resolution WSIs.
- To achieve this, we introduce a multi-magnification feature extraction strategy, which allows the model to learn visual representations across different spatial scales. This mechanism is further supported by auxiliary training objectives that guide the extraction of both fine-grained morphological patterns and broader tissue-level structures.
- Furthermore, we incorporate a prototype-based spatial modeling component, in which a set of prototype embeddings is learned to represent the most salient and recurring patterns across the entire slide. This prototype bank serves as a compact and informative summary of the global tissue context, enabling the model to reason about long-range spatial dependencies without the computational burden associated with dense pairwise patch interactions.

– We conduct comprehensive experiments to evaluate the performance of MMAP and compare it with existing approaches. Empirical results demonstrate the superiority of our method against state-of-the-arts.

2 Related Work

This section delves into existing studies pertinent to our research. From now on, we refer to the term "spot" as a predefined image patch within a WSI where gene expression is quantified, and use them interchangeably.

Deep features for WSIs. Advancements in deep learning have popularized feature extraction from pretrained networks like ResNet101 [22, 28], offering a strong foundation for pathology image analysis. Leveraging these features, Shao et al. [20] propose a graph embedding algorithm driven by tumor microenvironment interactions to represent image patches. Similarly, Chan et al. [4] introduce a heterogeneous graph learning approach to capture local structures in whole-slide image (WSI) pathology data. Recent developments in pathology foundation models are revolutionizing both general and medical AI, enabling the creation of versatile, general-purpose models that can be either frozen or fine-tuned to extract deep features from pathology images. For example, Lenz et al. [11] integrate features from multiple foundation models: UNI [5], CTransPath [24], Virchow2 [32], and H-optimus-0 [18] to generate comprehensive slide-level representations. Likewise, Song et al. [21] utilize UNI [5] to develop morphological slide-level representation sets. In addition, recent solutions [14, 15] employ parameter-efficient fine-tuning techniques, such as low-rank adaptation [10], to tailor general-purpose models for specific pathology tasks.

Spatial gene expression prediction from WSIs. We review key contributions in predicting spatial gene expression from whole-slide images (WSIs). ST-Net [8] employs a transfer learning approach, fine-tuning a DenseNet121 model [31], pretrained on ImageNet, to predict gene expression from histology images. Building on this, HisToGene [16] uses Vision Transformer to capture patch correlations across WSIs, enabling gene expression prediction with global context-aware features. Hist2ST [29] and TCGN [25] advance this further by incorporating neighborhood information through graph convolutional networks [7], emphasizing inter-spot relationships. In contrast, EGN [27] adopts exemplar learning to predict gene expression by selecting the most relevant exemplars from a spot within a WSI. However, this approach neglects local context around patches, compromising the balance between local and global feature integration.

3 Proposed Method

3.1 Problem Definition and Our Approach

Problem Definition. The goal of spatial gene expression (SGE) prediction is to estimate spatially resolved gene expression profiles directly from Hema-

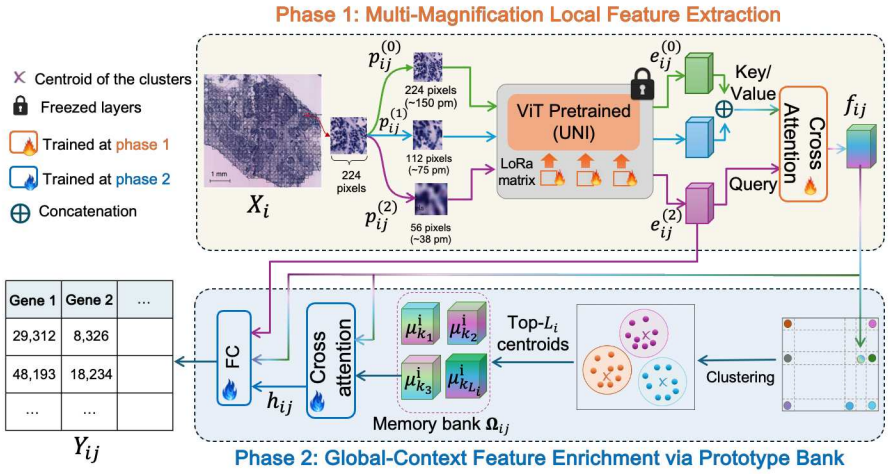

Fig. 1: Overview of the proposed framework, comprising two main phases: (1) Local feature extraction, and (2) global context-aware feature enrichment.

toxylin and Eosin (H&E)-stained WSIs, enabling in-silico molecular profiling without requiring costly spatial transcriptomics assays. Formally, given a WSI $X_i \in [0, 255]^{h \times w}$, the tissue section is partitioned into N_i non-overlapping image patches $\{P_{ij}\}_{j=1}^{N_i}$ of fixed size $p \times p$, each centered at the coordinates C_{ij} corresponding to spatial transcriptomic spot $Y_{ij} \in \mathbb{R}^g$, where g is the number of target genes. The task is to learn a regression model $\mathcal{F}_\theta$ that predicts

$$\hat{Y}_{ij} = \mathcal{F}_\theta(P_{ij}, C_{ij}), \tag{1}$$

such that $\hat{Y}_{ij}$ approximates Y_{ij} as closely as possible.

Our Solution. We introduce a unified framework that integrates multi-magnification local features with global prototypes to enhance gene expression prediction. This reformulates Eq. 1 as follows:

$$\mathcal{F}_\theta(Pij, C_{ij}) = \mathcal{F}^g_{\theta_2}\left(\mathcal{F}^\ell_{\theta_1}(Pij, C_{ij})\right),$$

where $\mathcal{F}^\ell_{\theta_1}$ extracts multi-magnification local features from patch P_{ij}, which are then used by $\mathcal{F}^g_{\theta_2}$ to construct a prototype bank for refining local features with global context.

Specifically, our approach operates in two phases. First, the patch-level feature extractor $\mathcal{F}^\ell_{\theta_1}$ learns discriminative morphological representations from image patches, encoding them into latent embeddings that capture key histological traits predictive of gene expression. In the second phase, a global context enhancement module $\mathcal{F}^g_{\theta_2}$ constructs a prototype bank and employs cross-attention

to model long-range dependencies across spatially distributed patches. This enriches local representations with global tissue context, enabling effective reasoning about both localized structures and broader WSI architecture. As shown in Fig. 1, this two-phase design balances expressive power and computational efficiency, fully leveraging the richness of whole-slide histopathology images while maintaining scalability and predictive accuracy.

3.2 Multi-magnification Local Feature Extraction

In the first phase of MMAP, we aim to learn robust patch-level representations by extracting features from histology patches at multiple magnification levels, inspired by pathologists' diagnostic workflow where varied zooms reveal complementary biological cues. For each patch, we generate sub-patches at $1/2$ and $1/4$ the original size (approximating $\times 10$ and $\times 20$ magnifications) via random cropping, process them through specialized embedding modules to capture magnification-specific features, and apply attention layers with auxiliary loss functions to focus on informative regions and enhance representation learning, as illustrated in Fig. 2.

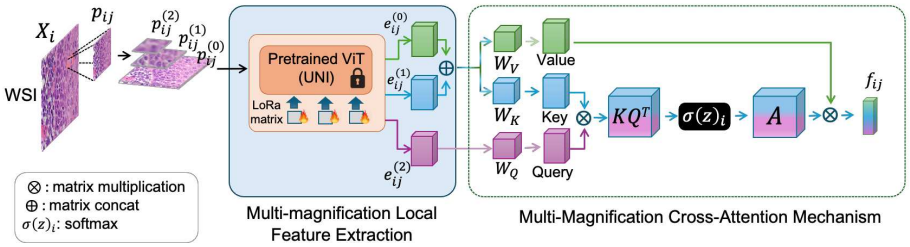

Fig. 2: Patch-level feature extraction with Multi-magnification enhancement.

Multi-Magnification Cross-Attention Mechanism. In this work, we propose a cross-magnification attention mechanism to dynamically model interactions between patches at different resolution levels. For each input patch P_{ij} of size $p \times p$ at the resolution $\times 5$, we generate two additional sub-patches at higher resolutions via random cropping:

$$P_{ij}^{(1)} \in \mathbb{R}^{\frac{p}{2} \times \frac{p}{2}}, \qquad P_{ij}^{(2)} \in \mathbb{R}^{\frac{p}{4} \times \frac{p}{4}}, \tag{2}$$

corresponding to approximate magnifications of $\times 10$ and $\times 20$, respectively. All patches are resized back to $p \times p$ for uniform processing, forming a multi-magnification input set:

$$P_{ij} = \left\{ \text{resize}(P_{ij}), \text{resize}(P_{ij}^{(1)}), \text{resize}(P_{ij}^{(2)}) \right\}. \tag{3}$$

These multi-magnification patches are independently processed by $\mathcal{F}^{\ell}_{\theta_1}$ to obtain magnification-specific embeddings:

$$\mathcal{E}_{ij} = \left\{ \mathbf{e}^{(0)}_{ij}, \mathbf{e}^{(1)}_{ij}, \mathbf{e}^{(2)}_{ij} \right\}, \tag{4}$$

where $\mathbf{e}^{s}_{ij} = \mathcal{F}^{\ell}_{\theta_1}(P^{s}_{ij})$. Here, s denotes the magnification level of the input patch. In this work, we fine-tune UNI [5], a foundation model pretrained on a large amount of histological images using LoRA adaption [10].

Using UNI-based encoder $\mathcal{F}^{\ell}_{\theta_1}$ on different magnification-specific inputs $\{P^{s}_{ij}\}$, $\forall s = \{1, 2, 3\}$, we generate sequences of τ vectors $\{z^{s}_{k}\}^{\tau}_{k=1}$, respectively, where the first tokens ([CLS] tokens) z^{s}_{0} represent the entire sequences. We construct an intermediate sequence $\{z^{0}_{0}, z^{1}_{1}, \ldots, z^{1}_{\tau}, z^{2}_{1}, \ldots, z^{2}_{\tau}\}$ and apply self-attention to enable the cell [CLS] token z^{0}_{0} to interact with features from the higher magnification levels, computing pairwise attention scores to focus on the most relevant information. Intuitively, this multi-magnification attention process aims to enrich representations by looking closer at each sub-region of the original given patch. The output of this process is a fused representation f^{0}_{0}, which combines z^{0}_{0} with magnification-specific features weighted by their attention scores, capturing complex interactions across patches and resolutions in histology images. Henceforth, for simplicity, we drop the subscript 0 of the [CLS] token and denote f_{ij} as the [CLS] vector of the input P_{ij}. f_{ij} is then fed into a linear layer to predict gene expressions.

Training Objectives. Annotation inconsistencies and the dropout phenomenon in gene expression typically introduce bias into the training data. To address this and ensure stability during training, we propose a contrastive magnification loss $\mathcal{L}_{\mathrm{mag}_1}$ to regularize f_{ij} and reduce dataset bias:

$$\mathcal{L}_{\mathrm{mag}_1} = 1 - \cos\left(f_{ij}, \mathbf{e}^{(0)}_{ij}\right), \tag{5}$$

where $\cos(\cdot)$ denotes the cosine similarity function. This loss encourages the original and magnification-enhanced representations of the same instance to align, ensuring stable training. For gene expression prediction, we employ a simple ℓ_2 loss function $\mathcal{L}_{\mathrm{ge}_1}$ to train the regression model. Together, the final loss of MMAP in the first stage is as follows:

$$\mathcal{L} = \mathcal{L}_{\mathrm{ge}_1} + \gamma_1 \mathcal{L}_{\mathrm{mag}_1}, \tag{6}$$

where γ_1 controls the impact of regularization losses.

3.3 Global-Context Feature Enrichment via Prototype Bank

After obtaining a robust patch-level feature extractor $\mathcal{F}^{\ell}_{\theta_1}$ from Phase 1, we freeze its parameters and use it as a feature inference backbone for global-context modeling. Specifically, for each WSI X_i, we process all patches $\{P_{ij}\}$ through phase 1's pipeline to extract two types of intermediate features: (1) the fused embeddings f_{ij} from the multi-magnification cross-attention module, and (2) the corresponding MLP regression outputs $\hat{Y}^{(1)}_{ij}$ from Phase 1.

Global Prototype Bank Construction. This section introduces a novel prototype bank, which serves as a condensed but rich representation set of the WSI. Technically, to capture global tissue context while avoiding full self-attention over all patches (which is computationally prohibitive), we perform K-means clustering on the set of fused embeddings $\{\mathbf{e}_{ij}^{\text{fused}}\}$ for each WSI X_i individually:

$$\{\mu_k^i\}_{k=1}^{K_i} = \text{KMeans}\left(\{f_{ij}\}_{j=1}^{N_i}\right), \tag{7}$$

where μ_k^i denotes the k-th cluster centroid for slide X_i, and K_i is adaptively selected based on the total number of patches in X_i to balance between over-sparsification and under-representation. For each patch P_{ij}, we retrieve its top-L_i most similar cluster centroids according to cosine similarity:

$$\Omega_{ij} = \text{TopL}_{L_i}\left(\{\mu_k^i\}, f_{ij}\right), \tag{8}$$

where $\Omega_{ij} = \left\{\mu_{k_1}^i, \ldots, \mu_{k_{L_i}}^i\right\}$ denotes our proposed prototype bank for patch P_{ij}. We refer to each embedding in Ω_{ij} as a global prototype due to its representativeness for the WSI.

We experiment with both fixed and adaptive retrieval strategies for L_i. In the adaptive setting, L_i is dynamically adjusted for each WSI based on the total number of clusters K_i to maintain a balance between capturing sufficient global diversity and avoiding overfitting to noisy centroids. As shown in Sec. 4, we empirically find that adaptive retrieval improves performance in case of highly variable tissue sizes.

Local-Global Cross-Attention Fusion. Given global prototype bank Ω_{ij}, we now aim to fuse information from local and global views to enrich the prediction process. Specifically, we design a global-context aware cross-attention module that integrates local patch features with retrieved global prototypes. Similar to the cross-attention process described in Sec. 3.2, for each patch P_{ij}, we produce a rich representation h_{ij}. In this attention process, f_{ij} serves as the queries and Ω_{ij} as keys and values. Since h_{ij} is now integrated with both local and global information, we linearly project h_{ij} to get another prediction for gene expression. The final prediction of MMAP can be formulated as:

$$\hat{Y}_{ij} = \left(\text{MLP}_1(\mathbf{e}_{ij}^{(2)}) + \text{MLP}_2(f_{ij}) + \text{MLP}_3(h_{ij})\right)/3, \tag{9}$$

where MLP denotes a linear projection layer.

Training Objectives. To ensure the prototype-enhanced features do not diverge, we similarly adopt the magnification loss $\mathcal{L}_{mag}$ in Sec. 3.2 as follows:

$$\mathcal{L}_{\text{mag}_2} = 1 - \cos\left(f_{ij}, h_{ij}\right). \tag{10}$$

A simple ℓ_2 loss function $\mathcal{L}_{\text{ge}_2}$ to train the regression model, leading the final loss of MMAP in the second stage:

$$\mathcal{L} = \mathcal{L}_{\text{ge}_2} + \gamma_2 \mathcal{L}_{\text{mag}_2}, \tag{11}$$

where γ_2 controls the impact of regularization losses.

4 Evaluation

4.1 Experimental Settings

Dataset. We conduct experiments on the publicly available HER2-positive breast cancer dataset [2], which includes 36 H&E-stained WSIs from 8 patients. For each tissue section, histology images, gene expression profiles, and spatial barcode coordinates are provided. We extract fixed-size image patches centered at sequencing spots and retain the top 1,000 highly variable genes (HVGs), further filtering genes expressed in fewer than 1,000 spots across the dataset. This preprocessing results in a total of 9,612 spots and 785 genes. The dataset is split at the slide level: 28 WSIs are used for training, and 8 WSIs with pathologist annotations are reserved for testing.

Baseline Methods. We compare our proposed method with four state-of-the-art baseline methods: ST-Net [8], DeepPT [9], HisToGene [16] and TCGN [25]. In particular, ST-Net and DeepPT both propose a CNN-based model to learn local features from the whole slide images; HisToGene utilizes a Transformer-based architecture to learn global relationships between patches of a whole slide image; TCGN attempts to combine local and global features while also using Graph Neural Networks (GNNs) to learn inter-spot relationships. These baselines cover different approaches in the field, ensuring the completeness of our experiments.

Evaluation Metrics. We evaluate our method's performance using three complementary metrics: Pearson Correlation Coefficient (PCC), Mean Squared Error (MSE), and Mean Absolute Error (MAE), all of which are widely adopted in performance testing [23]. MSE and MAE measure the degree of error between predicted and observed gene expression values by different normalization criteria. Both are used to provide a normalized measure of prediction accuracy. Lower error values indicate better prediction performance. Meanwhile, PCC measures the linear correlation between predicted and actual gene expression values across spatial locations. A higher PCC value indicates a stronger linear relationship, reflecting the accuracy of spatial gene expression prediction.

Implementation Details. All experiments are implemented using the PyTorch framework and trained on a server equipped with a single NVIDIA A100 GPU. We use the Adam optimizer with an initial learning rate of 1×10^{-5}, no weight

decay, and a batch size of 16. The learning rate is scheduled using Cosine Annealing over 50 epochs. Input images are resized to 112×112 and normalized using the ImageNet mean and standard deviation. Data augmentation includes random horizontal flip, rotation, and color jitter. In our proposed method, the number of clusters K in the K-means step is adaptively selected for each WSI based on its number of extracted patches. Specifically, K is dynamically adjusted within the range $[32, 80]$ to ensure a balanced trade-off between computational efficiency and fine-grained spatial representation. For gene expression values, we apply a $\log(1+x)$ transformation for normalization prior to training, which helps stabilize variance and reduce the effect of extreme values.

4.2 Experimental Results

Quantitative Evaluation. Table 1 presents a comprehensive comparison between our method (MMAP) and several state-of-the-art baselines on the gene expression prediction task. MMAP consistently outperforms all competitors across key evaluation metrics, demonstrating its superior capability in modeling spatial transcriptomic patterns from histopathology. In particular, MMAP achieves the highest PCC of **0.2619**, which is 3.5 times higher than the strongest baseline in this regard, HisToGene (<u>0.0753</u>). For error-based metrics, MMAP obtains the lowest MSE of **1.2439**, representing a 3.8% reduction compared to the second-best method, DeepPT (<u>1.2822</u>). Similarly, MMAP achieves a MAE of **0.8873**, yielding a 5.2% improvement over TCGN (<u>0.9396</u>). These results highlight the effectiveness of our proposed multi-magnification strategy in capturing fine-grained morphological and contextual cues that are predictive of underlying gene expression.

Table 1: Comparison of gene expression prediction performance on the test set. ↑/↓ means higher/lower values are better. The top and runner-up results are highlighted using **bold** and <u>underline</u>, respectively.

Method	ST-Net	HisToGene	DeepPT	TCGN	**MMAP (Ours)**
PCC ↑	0.0510	<u>0.0753</u>	0.0470	0.0515	**0.2619**
MAE ↓	0.9580	0.9664	0.9536	<u>0.9396</u>	**0.8873**
MSE ↓	1.4845	1.4562	<u>1.2822</u>	1.4858	**1.2439**

Qualitative Evaluation. In Figure 3, we present qualitative visualizations of predicted gene expression maps on the HER2+ dataset, where K-means clustering is applied to the outputs of each method to support spatial interpretation.

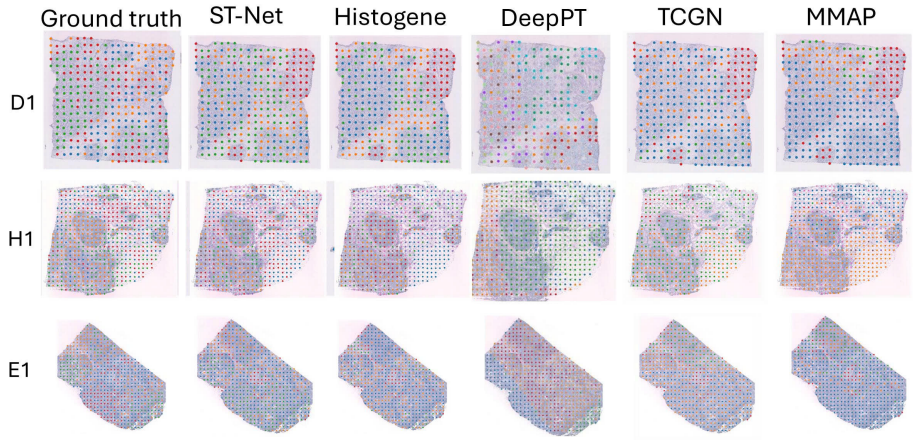

Fig. 3: Visualization of gene expression on the HER2+ dataset using prediction results obtained by all methods. The ground truth represents labels from the pathologist annotations. The illustrations for each figure are obtained by performing K-means clustering.

Ground-truth annotations derived from pathologist labels are used as references to assess spatial correspondence between predicted expression patterns and annotated tissue regions. MMAP predictions generally follow the structure of the annotated regions and delineate spatial domains with greater visual clarity, particularly in areas characterized by subtle morphological transitions. The clusters produced by MMAP tend to align with histological boundaries and exhibit coherent regional separation. In contrast, clusters generated by several baseline methods often appear spatially fragmented or less congruent with the underlying tissue architecture, resulting in overlapping or dispersed patterns across morphologically distinct regions. These observations highlight the ability of MMAP to produce structured, morphology-consistent predictions that are well-suited for downstream spatial analyses.

4.3 Ablation Study

Effect of Neighbor Selection Strategy. We conduct an ablation study to evaluate the impact of neighbor selection strategies during the second phase, where each patch of a WSI attends to its top-L most similar cluster centers. We compare fixed values of $L \in \{4, 8, 16, 32\}$ against an adaptive strategy, which selects the top 50% of clusters (i.e. $L = 0.5K$, where K is the number of clusters in the WSI) based on cosine similarity of the patch under consideration. As shown in Fig. 4, the adaptive approach consistently achieves higher PCC and lower MSE, MAE. We observe that increasing L generally leads to improved performance across all metrics, as a larger number of neighbors provide richer contextual information, helping the model better capture global structure and

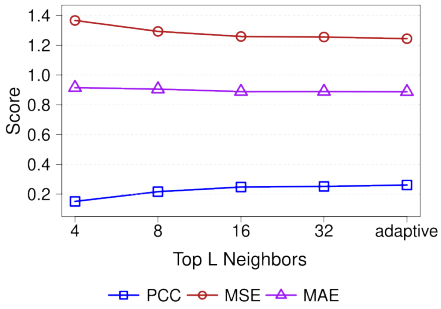

Fig. 4: Impact of the number of the selected prototypes L for a WSI. `adaptive` means $L = 0.5K$, where K is the number of clusters corresponding to the given WSI.

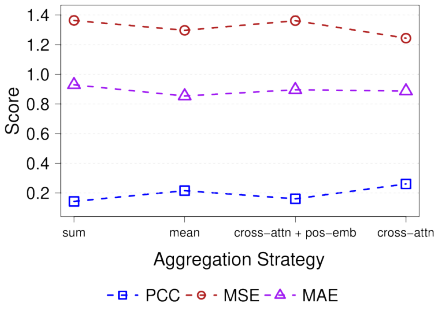

Fig. 5: Impact of four global context aggregation strategies: `mean` (element-wise mean), `sum` (element-wise sum), `cross-attn + pos-emb` (cross-attention with relative positional embeddings), and `cross-attn` (default cross-attention without positional encoding).

reduce noise in gene expression estimation. However, overly large or fixed values of L may still include irrelevant or redundant context, especially in slides with highly variable spatial distributions. In contrast, the adaptive strategy we propose offers a principled way to balance this trade-off by dynamically adjusting the number of neighbors based on the number of clusters per WSI. This flexibility enables MMAP to better adapt to local tissue complexity, yielding more consistent improvements in PCC, MSE, and MAE.

Ablation Study on Global Context Aggregation. We further investigate the role of the cross-attention module in aggregating information from neighboring cluster centers. Specifically, we replace the cross-attention block in phase two with three alternatives: (i) element-wise mean, (ii) element-wise sum over neighbor embeddings, and (iii) a variant that augments cross-attention with relative positional embeddings, computed as the coordinate offset between the center of the patch under consideration and the centers of its top-L neighbors. As shown in Fig. 5, the original cross-attention design consistently outperforms all alternatives across three evaluation metrics. This highlights its advantage in assigning adaptive relevance weights to neighboring regions based on learned contextual cues, rather than relying on fixed aggregation strategies. Notably, incorporating relative positional embeddings led to a consistent drop in performance. We attribute this to the possibility that explicit spatial encoding introduces inductive biases that overly constrain the attention mechanism. Since the cluster centers already encapsulate local structural information derived from spatially contiguous regions, enforcing additional position-based priors may hinder the model's flexibility in capturing semantically relevant but spatially distant patterns, ultimately impairing generalization in gene expression prediction.

5 Conclusion

In this study, we propose MMAP, a powerful two-phase deep learning model to predict gene expression profiles from histology images by combining both local information and global context. MMAP demonstrates strong performance in predicting spatial gene expression from H&E-stained histology images by effectively leveraging multi-magnification features and global context refinement. Unlike prior approaches that either overemphasize local features or rely heavily on computationally expensive global modeling, MMAP balances both through a prototype-based aggregation mechanism and a cross-magnification attention design. Our results show that MMAP consistently outperforms existing methods in prediction accuracy, confirming the utility of integrating hierarchical tissue information. These advantages allow MMAP to be more applicable in many real-world scenarios.

Acknowledgments. This study was funded by Hanoi University of Science and Technology (HUST) (Grant number T2024-TĐ-002)

Disclosure of Interests. The authors declare that they have no known competing interests.

Data availability. The spatial transcriptomics dataset that we used for this study can be found here at: https://github.com/almaan/her2st

Code availability. All source codes for baselines we used for comparison in our experiments are available at: https://github.com/lemonsoda174/MMAP-Baselines. The source code for MMAP is available at: https://github.com/pndh/MMAPattemp3.

References

1. Alon, S., Goodwin, D.R., Sinha, A., Wassie, A.T., Chen, F., Daugharthy, E.R., Bando, Y., Kajita, A., Xue, A.G., Marrett, K., Prior, R., Cui, Y., Payne, A.C., Yao, C.C., Suk, H.J., Wang, R., Yu, C.C.J., Tillberg, P., Reginato, P., Pak, N., Liu, S., Punthambaker, S., Iyer, E.P.R., Kohman, R.E., Miller, J.A., Lein, E.S., Lako, A., Cullen, N., Rodig, S., Helvie, K., Abravanel, D.L., Wagle, N., Johnson, B.E., Klughammer, J., Slyper, M., Waldman, J., Jané-Valbuena, J., Rozenblatt-Rosen, O., Regev, A., et al., I.C.: Expansion sequencing: Spatially precise in situ transcriptomics in intact biological systems. Science **371**(6528), eaax2656 (2021). https://doi.org/10.1126/science.aax2656, `https://www.science.org/doi/abs/10.1126/science.aax2656`
2. Andersson, A., Larsson, L., Stenbeck, L., Salmén, F., Ehinger, A., Wu, S.Z., Al-Eryani, G., Roden, D., Swarbrick, A., Borg, A., Frisén, J., Engblom, C., Lundeberg, J.: Spatial deconvolution of her2-positive breast cancer delineates tumor-associated cell type interactions. Nature Communications **12** (2021). https://doi.org/10.1038/s41467-021-26271-2, published 14 October 2021

3. Annaratone, L., Simonetti, M., Wernersson, E., Marchiò, C., Garnerone, S., Scalzo, M.S., Bienko, M., Chiarle, R., Sapino, A., Crosetto, N.: Quantification of her2 and estrogen receptor heterogeneity in breast cancer by single-molecule rna fluorescence in situ hybridization. Oncotarget **8**(12), 18680–18698 (Mar 2017). https://doi.org/10.18632/oncotarget.15727
4. Chan, T.H., Cendra, F.J., Ma, L., Yin, G., Yu, L.: Histopathology whole slide image analysis with heterogeneous graph representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15661–15670 (June 2023)
5. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F.K., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., Williams, M., Oldenburg, L., Weishaupt, L.L., Wang, J., Vaidya, A., Le, L.P., Gerber, G., Sahai, S., Williams, W., Mahmood, F.: Towards a general-purpose foundation model for computational pathology. Nature Medicine **30**(3), 850–862 (2024). https://doi.org/10.1038/s41591-024-02857-3, epub 2024 Mar 19
6. 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)
7. Corso, G., Stark, H., Jegelka, S., Jaakkola, T., Barzilay, R.: Graph neural networks. Nature Reviews Methods Primers **4**(1), 17 (2024). https://doi.org/10.1038/s43586-024-00294-7, https://doi.org/10.1038/s43586-024-00294-7
8. 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 (Aug 2020)
9. Hoang, D.T., Dinstag, G., Shulman, E.D., Hermida, L.C., BenZvi, D.S., Elis, E., Caley, K., Sammut, S., Sinha, S., Sinha, N., Dampier, C.H., Stossel, C., Patil, T., Rajan, A., Lassoued, W., Strauss, J., Bailey, S., Allen, C., Redman, J., Beker, T., Jiang, P., Golan, T., Wilkinson, S., Sowalsky, A.G., Pine, S.R., Caldas, C., Gulley, J.L., Aldape, K., Aharonov, R., Stone, E.A., Ruppin, E.: A deep-learning framework to predict cancer treatment response from histopathology images through imputed transcriptomics. Nature Cancer **5**(9), 1305–1317 (2024). https://doi.org/10.1038/s43018-024-00793-2, epub 2024 Jul 3
10. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, e.a.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
11. Lenz, T., Neidlinger, P., Ligero, M., Wölflein, G., van Treeck, M., Kather, J.N.: Unsupervised foundation model-agnostic slide-level representation learning. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR). pp. 30807–30817 (June 2025)
12. Lugmayr, W., Kotov, V., Goessweiner-Mohr, N., Wald, J., DiMaio, F., Marlovits, T.C.: Starmap: a user-friendly workflow for rosetta-driven molecular structure refinement. Nature Protocols **18**(1), 239–264 (2023). https://doi.org/10.1038/s41596-022-00757-9, https://doi.org/10.1038/s41596-022-00757-9
13. Nagendran, M., Sapida, J., Arthur, J., Yin, Y., Tuncer, S.D., Anaparthy, N., Gupta, A., Serra, M., Patterson, D., Tentori, A.: 1457 visium hd enables spatially resolved, single-cell scale resolution mapping of ffpe human breast cancer tissue. Journal for ImmunoTherapy of Cancer **11**(Suppl 1) (2023). https://doi.org/10.1136/jitc-

2023-SITC2023.1457, https://jitcsite-bmj.vercel.app/content/11/Suppl_1/A1620, published online: November 2, 2023
14. Nguyen, M.D., Huy Pham, N.D., Nguyen, P.L., Do, M.N.: A semi-supervised learning framework with cross-magnification attention for glioma mitosis classification. In: 2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI). pp. 1–4 (2025). https://doi.org/10.1109/ISBI60581.2025.10981240
15. Nguyen, M.D., Nguyen, D.T., Nguyen, T.V., Yamada, H., Pham, H.H., Nguyen, P.L.: Bridging classification and segmentation in osteosarcoma assessment via foundation and discrete diffusion models (2025), https://arxiv.org/abs/2501.01932
16. Pang, M., Su, K., Li, M.: Leveraging information in spatial transcriptomics to predict super-resolution gene expression from histology images in tumors. bioRxiv (2021). https://doi.org/10.1101/2021.11.28.470212, https://www.biorxiv.org/content/early/2021/11/28/2021.11.28.470212
17. Pang, M., Su, K., Li, M.: Leveraging information in spatial transcriptomics to predict super-resolution gene expression from histology images in tumors. bioRxiv (2021). https://doi.org/10.1101/2021.11.28.470212, https://www.biorxiv.org/content/early/2021/11/28/2021.11.28.470212
18. Saillard, C., Jenatton, R., Llinares-López, F., Mariet, Z., Cahané, D., Durand, E., Vert, J.P.: H-optimus-0 (2024), https://github.com/bioptimus/releases/tree/main/models/h-optimus/v0
19. Shah, S., Takei, Y., Zhou, W., Lubeck, E., Yun, J., Eng, C.H.L., Koulena, N., Cronin, C., Karp, C., Liaw, E.J., Amin, M., Cai, L.: Dynamics and spatial genomics of the nascent transcriptome by intron seqfish. Cell **174**(2), 363–376.e16 (2018). https://doi.org/10.1016/j.cell.2018.05.035, https://doi.org/10.1016/j.cell.2018.05.035
20. Shao, W., Shi, Y., Zhang, D., Zhou, J., Wan, P.: Tumor Micro-Environment Interactions Guided Graph Learning for Survival Analysis of Human Cancers from Whole-Slide Pathological Images . In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11694–11703. IEEE Computer Society, Los Alamitos, CA, USA (Jun 2024). https://doi.org/10.1109/CVPR52733.2024.01111, https://doi.ieeecomputersociety.org/10.1109/CVPR52733.2024.01111
21. Song, A.H., Chen, R.J., Ding, T., Williamson, D.F., Jaume, G., Mahmood, F.: Morphological prototyping for unsupervised slide representation learning in computational pathology. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
22. Srinidhi, C.L., Ciga, O., Martel, A.L.: Deep neural network models for computational histopathology: A survey. Med Image Anal **67**, 101813 (Sep 2020)
23. Wang, C., Chan, A.S., Fu, X., Ghazanfar, S., Kim, J., Patrick, E., Yang, J., et al.: Benchmarking the translational potential of spatial gene expression prediction from histology. Nature Communications **16** (2025). https://doi.org/10.1038/s41467-025-56618-y, open access; Published 11 February 2025
24. Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., Han, X.: Transformer-based unsupervised contrastive learning for histopathological image classification. Medical Image Analysis **81**, 102559 (2022). https://doi.org/https://doi.org/10.1016/j.media.2022.102559, https://www.sciencedirect.com/science/article/pii/S1361841522002043
25. 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). https://doi.org/https://doi.org/10.1016/j.media.2023.103040, https://www.sciencedirect.com/science/article/pii/S1361841523003006

26. Xie, R., Pang, K., Chung, S., Perciani, C., MacParland, S., Wang, B., Bader, G.: Spatially resolved gene expression prediction from histology images via bi-modal contrastive learning. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 70626–70637. Curran Associates, Inc. (2023), https://proceedings.neurips.cc/paper_files/paper/2023/file/df656d6ed77b565e8dcdfbf568aead0a-Paper-Conference.pdf
27. Yang, Y., Hossain, M.Z., Stone, E.A., Rahman, S.: 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 (WACV). pp. 5039–5048 (January 2023)
28. Yao, J., Zhu, X., Jonnagaddala, J., Hawkins, N., Huang, J.: Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. Medical Image Analysis p. 101789 (July 2020). https://doi.org/10.1016/j.media.2020.101789, https://linkinghub.elsevier.com/retrieve/pii/S1361841520301535
29. Zeng, Y., Wei, Z., Yu, W., Yin, R., Li, B., Tang, Z., Lu, Y., Yang, Y.: Spatial transcriptomics prediction from histology jointly through transformer and graph neural networks. bioRxiv (2022). https://doi.org/10.1101/2022.04.25.489397, https://www.biorxiv.org/content/early/2022/04/26/2022.04.25.489397
30. Zhang, M., Eichhorn, S.W., Zingg, B., Yao, Z., Cotter, K., Zeng, H., Dong, H., Zhuang, X.: Spatially resolved cell atlas of the mouse primary motor cortex by merfish. Nature **598**(7879), 137–143 (2021). https://doi.org/10.1038/s41586-021-03705-x, https://doi.org/10.1038/s41586-021-03705-x
31. Zhu, Y., Newsam, S.: Densenet for dense flow. In: 2017 IEEE International Conference on Image Processing (ICIP). pp. 790–794 (2017). https://doi.org/10.1109/ICIP.2017.8296389
32. Zimmermann, E., Vorontsov, E., Viret, J., Casson, A., Zelechowski, M., Shaikovski, G., Tenenholtz, N., Hall, J., Klimstra, D., Yousfi, R., Fuchs, T., Fusi, N., Liu, S., Severson, K.: Virchow2: Scaling self-supervised mixed magnification models in pathology (2024), https://arxiv.org/abs/2408.00738