M2OST: Many-to-one Regression for Predicting Spatial Transcriptomics from Digital Pathology Images

Hongyi Wang[1], Xiuju Du[2], Jing Liu[2], Shuyi Ouyang[1], Yen-Wei Chen[3*], Lanfen Lin[1*]

[1]Zhejiang University, Hangzhou, China
[2]Zhejiang Lab, Hangzhou, China
[3]Ritsumeikan University, Osaka, Japan
chen@is.ritsumei.ac.jp, llf@zju.edu.cn

Abstract

The advancement of Spatial Transcriptomics (ST) has facilitated the spatially-aware profiling of gene expressions based on histopathology images. Although ST data offers valuable insights into the micro-environment of tumors, its acquisition cost remains expensive. Therefore, directly predicting the ST expressions from digital pathology images is desired. Current methods usually adopt existing regression backbones along with patch-sampling for this task, which ignores the inherent multi-scale information embedded in the pyramidal data structure of digital pathology images, and wastes the inter-spot visual information crucial for accurate gene expression prediction. To address these limitations, we propose M2OST, a many-to-one regression Transformer that can accommodate the hierarchical structure of the pathology images via a decoupled multi-scale feature extractor. Unlike traditional models that are trained with one-to-one image-label pairs, M2OST uses multiple images from different levels of the digital pathology image to jointly predict the gene expressions in their common corresponding spot. Built upon our many-to-one scheme, M2OST can be easily scaled to fit different numbers of inputs, and its network structure inherently incorporates nearby inter-spot features, enhancing regression performance. We have tested M2OST on three public ST datasets and the experimental results show that M2OST can achieve state-of-the-art performance with fewer parameters and floating-point operations (FLOPs).

Code — https://github.com/Dootmaan/M2OST

Introduction

Digital pathology images, as a kind of Whole Slide Images (WSIs), have witnessed widespread utilization in research nowadays, as they can be more easily stored and analyzed compared to traditional glass slides (Niazi, Parwani, and Gurcan 2019). However, besides the spatial organization of cells presented in these pathology images, the spatial variance of gene expressions is also very important for unraveling the intricate transcriptional architecture of multicellular organisms (Rao et al. 2021; Tian, Chen, and Macosko 2023; Cang et al. 2023). As the extended technologies of single-cell RNA sequencing (Kolodziejczyk et al. 2015; Mrabah et al. 2023), ST technologies have been developed recently, facilitating such spatially-aware profiling of gene expressions within tissues (Rodriques et al. 2019; Lee et al. 2021; Bressan, Battistoni, and Hannon 2023).

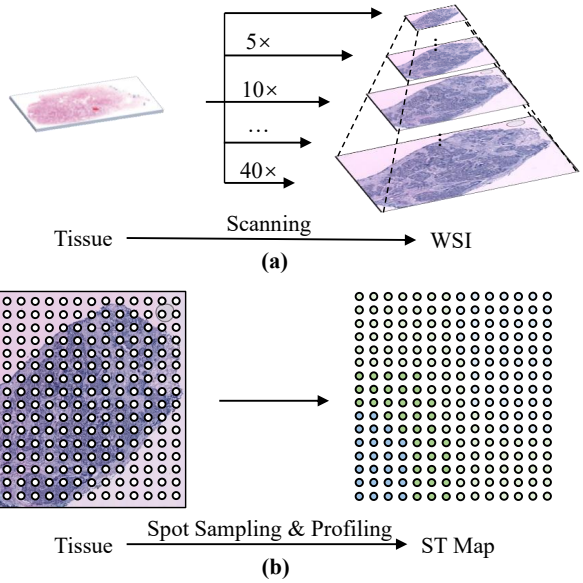

Figure 1: (a) WSIs are obtained by scanning the glass slide tissues at different magnifications, resulting in a multi-scale pyramid data structure. (b) ST maps are generated by sampling spots on the glass slide tissues, followed by comprehensive profiling of gene expressions within each spot.

A detailed illustration of the acquisition process of WSIs and ST maps is presented in Figure 1. As shown, WSIs are obtained by scanning the glass slide tissues at various magnification factors, resulting in a multi-scale hierarchical data structure (Ryu et al. 2023). Correspondingly, ST maps are obtained by firstly sampling spots with a fixed interval on the glass slide tissues. Each spot contains two to dozens of cells depending on different ST technologies (Song and Su 2021). Subsequently, the accumulated gene expressions of the cells within the spots are profiled, forming a spatial gene expression map. Such gene expression maps can be used along with their corresponding WSIs for multi-modal computational pathology analysis, leading to higher perfor-

*Corresponding authors.

mance in tasks such as cancer sub-typing and prognosis prediction (Hoang et al. 2022). However, despite their rapid evolution, ST technologies have yet to find widespread application in pathological analysis, primarily due to the expensive costs (Pang, Su, and Li 2021). In contrast, WSIs are more economical and accessible as they are routinely generated in clinics (Pang, Su, and Li 2021). Consequently, there is a growing imperative to directly generate ST maps from WSIs at a low cost through deep learning methods (Levy-Jurgenson et al. 2020; Weitz et al. 2021).

Current approaches typically treat the ST prediction problem as a conventional regression problem (He et al. 2020; Monjo et al. 2022), where the network is fed with a WSI patch as input and produces the cumulative gene expression intensities of the cells within the corresponding patch area. In this paradigm, the methods are trained with single-level image-label pairs just like standard regression tasks. This makes them only able to model the relationship between the gene expressions and the images of the maximum magnification, wasting the multi-scale information inherent in WSIs. From a bionic perspective, pathologists often zoom in and zoom out frequently when analyzing WSIs, as each level of WSIs encapsulates distinct morphological information that can be useful for ST predictions (Chen et al. 2022; Yarlagadda, Massagué, and Leslie 2023). For instance, cell-level images can facilitate the evaluation of gene expressions based on cell types, while higher-level images can offer regional morphologies that help determine overall gene intensities. Hence, we propose to conceptualize the ST prediction as a many-to-one modeling problem, in which case multiple images from different levels of WSI are leveraged to jointly predict the gene expressions within the spots. As we notice that the absolute field of view of the microscope will not change during the zooming operations performed by pathologists, we also biomimetically employ a fixed patch size for the pathology patches from different WSI levels in our regression model. In this case, higher-level image patches naturally have a larger receptive field, and thus is able to include more supporting features around the ST spot, compensating for the destroyed cell features on the patch edges during the patch cropping procedure (Chung et al. 2024).

Many-to-one-based modeling aims to learn a mapping function from a variable number of inputs to one single output. These multiple inputs can have different shapes or lack semantic alignment, with the goal being to find their common mapping target. It can be used for many tasks, such as multi-phase radiology image analysis (Hu et al. 2023) and label assignment problem (Wei et al. 2023). Our many-to-one scheme differs from conventional multi-scale methods by offering a structure that can easily scale to accommodate different numbers and different shapes of inputs. For instance, while we primarily present our model in a three-to-one structure, it can be easily adjusted to two-to-one or four-to-one scenarios by removing or adding streams in the pipeline, making it suitable for different WSI scanning technologies. Additionally, during training, model parameters can be partially updated when some levels of inputs are missing, as the model parameters are highly decoupled across the multiple inputs.

Based on this idea, we propose M2OST, a many-to-one-based regression Transformer designed to leverage pathology images at various levels to jointly predict the gene expressions. By incorporating the inter-spot visual information and the multi-scale features within the WSIs, M2OST exhibits the capability to generate more accurate ST maps. Moreover, to optimize the computational efficiency, we further introduce Intra-Level Token Mixing Module (ITMM), Cross-Level Token Mixing Module (CTMM), and Cross-Level Channel Mixing Module (CCMM) to decouple the many-to-one multi-scale feature extraction process into intra-scale representation learning and cross-scale feature interaction processes, which greatly reduces the computational cost without compromising model performance. In summary, our contributions are:

1. We propose to conceptualize the ST prediction problem as a many-to-one modeling problem, leveraging the multi-scale information and inter-spot features embedded in the hierarchically structured WSIs for joint prediction of the ST maps.
2. We propose M2OST, a flexible regression Transformer crafted to model many-to-one relationships for ST prediction. Its unique design makes M2OST suitable for different many-to-one scenarios, and is robust to input sets with various sequence lengths.
3. In M2OST, we propose to decouple the multi-scale feature extraction process into intra-scale feature extraction and cross-scale feature extraction, which significantly improves the computational efficiency without compromising model performance.
4. We have conducted thorough experiments on the proposed M2OST method, and have proved its effectiveness with three public ST datasets.

Related Works

The prediction of ST maps from WSIs has garnered sustained attention since the inception of ST technologies. ST-Net (He et al. 2020) is the first work that attempts to tackle this problem. ST-Net employs a convolutional neural network (CNN) with dense residual connections (He et al. 2016; Huang et al. 2017) to predict patch-wise gene expressions. By sequentially processing the patches in a WSI, ST-Net can eventually generate a complete ST map. Similarly, DeepSpaCE (Monjo et al. 2022) adopts a VGG-16 (Simonyan and Zisserman 2015) based CNN for such patch-level ST prediction, and it introduces semi-supervised learning techniques to augment the training sample pool. More recently, BLEEP (Xie et al. 2024) introduced a contrastive learning approach to align WSI patch features with ST spot embeddings, using K-Nearest Neighbors during the inference stage to mitigate the batch effect in biomedical datasets.

Although these classic CNN backbones have demonstrated considerable success in various vision tasks, their performance has been eclipsed by the advancements achieved with Transformer-based models (Ding et al. 2023). HisToGene (Pang, Su, and Li 2021) was the first method proposed to leverage vision Transformers (Dosovitskiy et al.

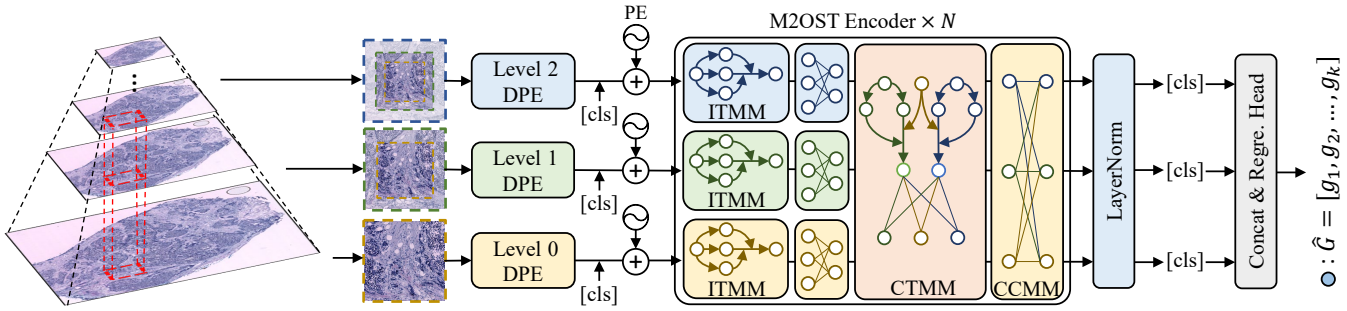

Figure 2: A schematic view of the proposed M2OST. Three patch sequences from different WSI levels are fed into the model to jointly predict the gene expressions in the corresponding spot. PE denotes the fully learnable positional embedding in the figure.

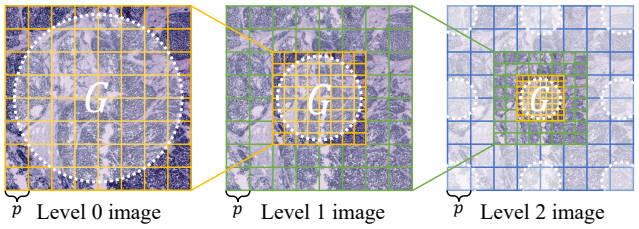

Figure 3: DPE used in M2OST. The circle area of G indicates the target ST spot.

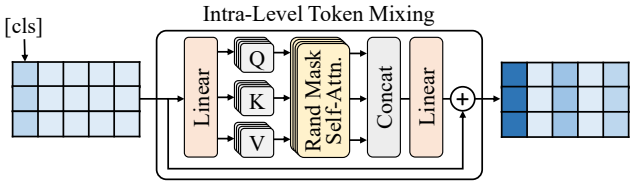

Figure 4: The network structure of ITMM. This module needs to be applied to each level's sequence separately.

2020) for predicting ST maps. Diverging from the approach of ST-Net and DeepSpaCE, which predict one spot at a time, HisToGene proposes to predict the entire ST map at a time. HisToGene takes the sequenced patches in a WSI as network input and employs the Self-Attention mechanism (Vaswani et al. 2017) to model the inter-correlations between these patches. Despite the efficiency gained from this slide-level scheme, the performance of HisToGene is constrained by the use of a relatively small ViT backbone, driven by computational limitations.

Following the path of HisToGene, Hist2ST (Zeng et al. 2022) was then proposed. Combining CNNs, Transformers, and Graph Neural Networks (Hamilton, Ying, and Leskovec 2017), Hist2ST strives to capture more intricate long-range dependencies. Like HisToGene, Hist2ST is also a slide-level method that uses the patch sequence as input to directly generate the gene expressions of all spots in an ST map. However, the complexity of its model structure results in considerable FLOPs and model size, elevating the risk of overfitting.

Contrary to the prevalent belief in the necessity of inter-spot correlations for predicting ST maps, iStar(Zhang et al. 2024) argues that gene expressions within a spot are logically related only to its corresponding patch area, thus reverting to a spot-level training scheme. It adopts HIPT(Chen et al. 2022), a hierarchical Vision Transformer pre-trained on large-scale WSI datasets for non-trainable slide-level feature extraction, and utilizes a simple MLP to fit the mapping relation from the feature maps to the ST spots, achieving state-of-the-art performance. However, as the feature extraction stage of iStar is unlearnable, it still leaves space for performance improvements. Building on this insight, in our proposed M2OST, we also adhere to the patch-level scheme, predicting a single spot at a time to ensure the independence and accuracy of each prediction.

Methodology

Problem Formulation

In M2OST, we use I_0, I_1, and $I_2 \in R^{3\times H\times W}$ to represent the three input images of different levels, where I_i denotes the pathology image patch from level i, and H, W represents the image height and width, respectively. The observed gene expressions in each spot are denoted as $G = \{g_1, g_2, ..., g_k\}$, where k is the total number of genes. The goal is to minimize the mean squared error (MSE) between $\hat{G} = \text{M2OST}(\{I_0, I_1, I_2\}|\theta_0, \theta_1, \theta_2)$ and G by optimizing the network parameters θ_0,θ_1, and θ_2 of each stream.

Overview of M2OST

A schematic view of the proposed M2OST is presented in Figure 2. Upon receiving the multi-scale pathology image patches from three different levels, M2OST initially sends them into our proposed Deformable Patch Embedding (DPE) layers to realize adaptive token generation. After appending $[cls]$ token to each sequence, intra-scale representation learning within each sequence is first performed using ITMM. Then, CTMM is introduced to facilitate cross-scale information exchange between the different inputs, followed by CCMM mixing the channels in a squeeze-and-excitation way. This multi-scale feature extraction module is termed the M2OST Encoder and is iterated N times within M2OST.

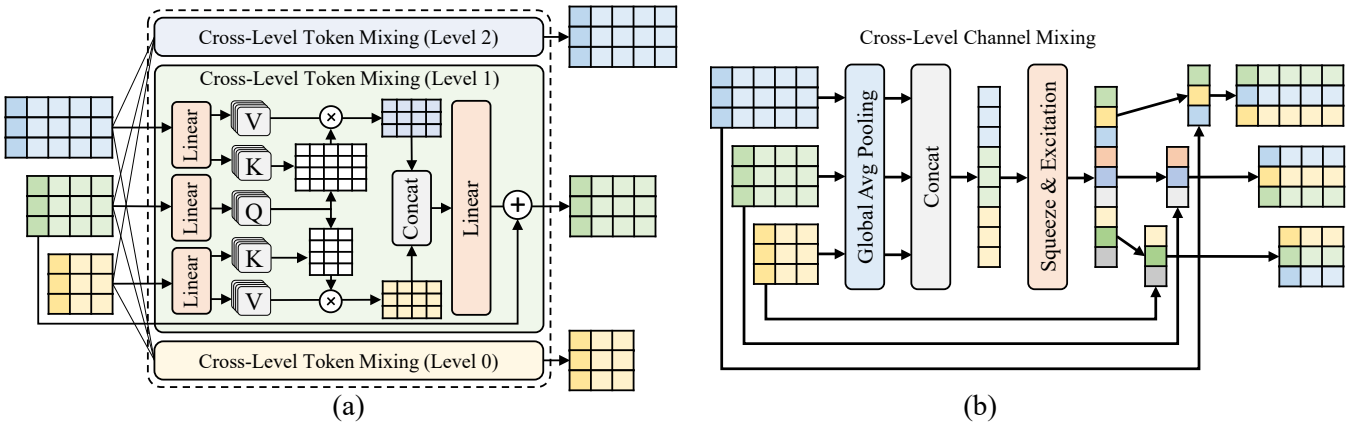

Figure 5: (a) The network structure of CTMM. (b) The network structure of CCMM.

Finally, the three $[cls]$ tokens are concatenated to be fed into the linear regression head for the ST spot prediction.

Deformable Patch Embedding

Although high-level pathology patches provide more nearby visual information of the target ST spot, the central image area that directly maps to the target spot should still be primarily focused on. To emphasize the in-spot features during many-to-one modeling, we introduce DPE to generate fine-grained in-spot tokens and coarse-grained surrounding tokens. As shown in Fig 3, apart from using patch size p on I_0, I_1, and I_2 to generate the basic tokens in a weight-sharing manner, DPE also adopts $\frac{p}{2}$ and $\frac{p}{4}$ patch sizes on the higher-level pathology images to ensure the central area of the patches receives the most attention. Eventually, DPE converts the three input images into $L \times C$, $2L \times C$, and $3L \times C$ sequences S_0, S_1, S_2, where $L = \frac{HW}{p^2}$ represents the sequence length and C denotes the number of channels after embedding.

Intra-Level Token Mixing

After patch embedding, a $[cls]$ token is appended to the beginning of each sequence. Then, a fully learnable positional embedding is added to each sequence to encode the positional information for the multi-scale tokens generated by DPE. After that, each level's sequence data undergoes processing through ITMM to extract the intra-level features, whose structure is mainly based on ViT (Dosovitskiy et al. 2020), as depicted in Figure 4. ITMM uses the Random Mask Self-Attention trick (Zehui et al. 2019) to enhance its generalization ability.

Cross-Level Token Mixing

After the intra-level feature extraction, the acquired representations are amalgamated into CTMM for cross-level information exchange. Given that the sequence lengths of S_0, S_1, and S_2 are different, these data can not be directly fused together for simple information exchange. In the meantime, to most possibly retain the independence of the network parameters θ_0, θ_1, and θ_2, CTMM introduces fully-connected cross-level attention to realize the goal, as illustrated in Fig 5(a).

Let M represents the total number of input sequences, and K_i, Q_i, V_i denotes the hidden representation extracted from S_i, CTMM can be mathematically defined as:

$$CTMM(S_i) = \sum_{0 \le j < M}^{j \neq i} \omega_j^i \cdot \sigma(\frac{Q_i K_j^T}{\sqrt{d_K}}) V_j, \qquad (1)$$

where σ represents the Softmax operation, and ω_j^i is the learnable weights indicating how much the j-th level information contributes to the i-th level CTMM output. Finally, M sequences are obtained with their shape unchanged, and subsequently sent into CCMM for further channel mixing.

Cross-Level Channel Mixing

CCMM is used to explicitly facilitate the channel interaction between multi-level sequences. Since the input sequences are still of different lengths, we design a length-insensitive channel mixing method for CCMM to address this issue, which is presented in Fig 5(b). Inspired by the squeeze-and-excitation operation in (Fu et al. 2019), we first use global average pooling for each sequence to compress their sequential information into one token. Then, we combine these tokens from different levels together and use a squeeze-and-excitation operation to obtain the cross-level channel attention scores. After that, the scores are split and multiplied back to their respective input sequences, leading to channel-level cross-scale information exchange.

In summary, as afore-presented, every module of M2OST is insensitive to sequence length, and can be easily scaled to handle different numbers of input by removing or adding streams to the pipeline.

Experiments

Datasets and Metrics

In our experiments, we utilized three public datasets to evaluate the performance of the proposed M2OST model.

The first one is the human breast cancer (HBC) dataset (Stenbeck et al. 2021). This dataset contains 30,612 spots in

DPE	ITMM	CTMM	CCMM	PCC (%)			Param# (M)	FLOPs (G)
				HBC	HER2+	cSCC		
✓	✓	✓	✓	**48.07**	**44.17**	**50.50**	**6.81**	2.24
	✓	✓	✓	47.13	43.10	49.35	7.76	**1.23**
		✓	✓	47.03	42.99	49.48	12.88	2.16
			✓	46.34	42.57	48.81	10.66	1.72
				46.12	42.55	48.66	10.66	2.07

Table 1: Ablation study results based on substituting components of M2OST into others.

Input			HBC		HER2+		cSCC	
Lvl 0	Lvl 1	Lvl 2	PCC(%)	RMSE	PCC(%)	RMSE	PCC(%)	RMSE
✓			46.92	3.17	43.12	3.06	49.31	3.60
	✓		45.23	3.18	42.56	3.11	48.27	3.81
		✓	41.04	3.25	40.01	3.21	45.29	3.89
✓	✓		47.32	3.16	43.31	3.05	50.02	3.47
✓		✓	46.94	3.17	42.98	3.10	49.73	3.61
	✓	✓	45.62	3.20	42.67	3.10	49.11	3.80
✓	✓	✓	**48.07**	**3.16**	**44.17**	**2.87**	**50.50**	**3.45**

Table 2: Ablation study on the input combinations of M2OST.

68 WSIs, and each spot has up to 26,949 distinct genes. The spots in this dataset exhibit a diameter of 100 µm, arranged in a grid with a center-to-center distance of 200 µm.

The second dataset is the human HER2-positive breast tumor dataset (Andersson et al. 2021). This dataset consists of 36 pathology images and 13,594 spots, and each spot contains 15,045 recorded gene expressions. Similar to the previous dataset, the ST data in this dataset also features a 200µm center-to-center distance between each captured spot with the diameter of each spot also being 100µm.

The third dataset is the human cutaneous squamous cell carcinoma (cSCC) dataset (Ji et al. 2020), which includes 12 WSIs and 8,671 spots. Each spots in this dataset have 16,959 genes profiled. All the spots have a diameter of 110µm and are arranged in a centered rectangular lattice pattern with a center-to-center distance of 150µm.

We employ the mean values of Pearson Correlation Coefficients (PCC) and Root Mean Squared Error (RMSE) of the spots to evaluate the regression accuracy. Mathematically, PCC can be described as:

$$PCC = \frac{Cov(G, \hat{G})}{Var(G) \cdot Var(\hat{G})}, \tag{2}$$

where $Cov(\cdot)$ is the covariance, $Var(\cdot)$ is the variance, G is the ground truth gene expressions of a spot, $\hat{G}$ is the corresponding predicted result.

Implementation Details

Given the inherently sparse nature of the ST map, we filter out less-variable genes in each dataset based on the criteria outlined in (He et al. 2020), eventually preserving 250 spatially variable genes per dataset for training. As for the preprocessing procedures, they are also kept identical to those described in (He et al. 2020). Specifically, we normalize the gene expression counts for each spot by dividing them by the sum of expressions within that spot, then multiplying the result by a scale factor of 1,000,000. The normalized values are subsequently transformed using the natural logarithm, calculated as $\log(1 + x)$, where x is the normalized count.

For all datasets, we use a patch size of 224×224 (which covers around 110µm×110µm in the pathology image) for each spot on level 0 pathology image, and the patch size p is set to 16 accordingly. In each dataset, 60% of the WSIs and their corresponding ST maps are used for training, 10% for validation, and the remaining 30% for testing. All the methods are trained with Adam (Kingma and Ba 2015) optimizer with a learning rate of 1e-4 for 100 epochs. Batch size is 96 for patch-level methods and 1 for slide-level methods. The hyper-parameters of M2OST are the model width, model depth, and the number of heads in self-attention. The three hyper-parameters were tuned following the goal of surpassing other methods with minimal model size. Specifically, the M2OST Encoder is repeated 4 times (i.e, model depth), the embedding channel is 192 (i.e., model width), and the number of head for the self-attention operation in ITMM is set to be 3. A larger model size can lead to even better ST regression performance but the computational cost will also be higher. All the methods are trained on two Nvidia RTX A6000 (48G) GPUs.

Ablation Study

Study on the M2OST Model Structure. To verify the effectiveness and efficiency of M2OST, we have conducted a thorough ablation study on its network structure, of which the experimental results are presented in Table 1. We begin by replacing DPE with ordinary patch embedding layers, which leads to a notable decrease in PCC of all three datasets, namely 0.94%, 1.07%, and 1.15%. Although the FLOPs dropped due to the reduced input sequence length, the parameter counts increased because of the absence of the weight-sharing mechanism used in DPE. Such experimental results prove the effectiveness of the adaptive patch embedding in DPE.

Then, we substitute the three ITMMs into one unified Self-Attention to directly process the concatenated sequences (the three sequences are of the same length without DPE, so they can be directly concatenated), destroying the decoupled design in M2OST. It is observed that the parameter count dramatically increased, but the model performance did not benefit from it, which validates the efficiency of using ITMM to decouple the multi-scale feature extraction process in M2OST. We further remove CTMM from M2OST, using simple concatenation for cross-level feature fusion. This time, the parameter count did not drop much, while the performance suffered a further decline. This indicates that CTMM is necessary for processing such many-to-one modeling problems, where each sequence may contain different semantic information that cannot be fused by simple concatenation. We have also tried using summation to

Methods	HBC		HER2+		cSCC		Parameter Count (M)	FLOPs (G)
	PCC(%)	RMSE	PCC(%)	RMSE	PCC(%)	RMSE		
ResNet50 (He et al. 2016)	47.10	3.17	43.33	3.04	49.34	3.60	24.02	4.11
ViT-B/16 (Dosovitskiy et al. 2020)	46.67	3.17	43.78	3.09	49.01	3.77	57.45	11.27
Swin-T (Liu et al. 2021)	44.52	3.29	37.67	3.57	48.83	3.74	19.02	2.96
ConvNeXt-T (Liu et al. 2022)	47.25	3.16	43.56	3.07	<u>50.08</u>	<u>3.49</u>	27.99	4.46
CrossViT (Chen, Fan, and Panda 2021)	47.46	3.16	43.90	3.04	49.51	3.55	26.27	4.85
DeepSpaCE (Monjo et al. 2022)	46.01	3.19	42.57	3.17	48.99	3.73	135.29	15.48
ST-Net (He et al. 2020)	<u>47.78</u>	<u>3.16</u>	43.01	3.07	49.37	3.58	<u>7.21</u>	<u>2.87</u>
HisToGene (Pang, Su, and Li 2021)	44.76	3.20	36.97	3.62	45.71	3.93	187.99	135.07
Hist2ST (Zeng et al. 2022)	45.00	3.18	40.02	3.06	46.71	3.88	675.50	1063.23
BLEEP (Xie et al. 2024)	47.02	3.17	43.53	3.05	49.60	3.59	24.18	4.19
HIPT/iStar (Chen et al. 2022; Zhang et al. 2024)	47.60	3.16	<u>43.92</u>	<u>3.01</u>	49.73	3.52	24.59	5.13
M2OST (Ours)	**48.07**	**3.16**	**44.17**	**2.87**	**50.50**	**3.45**	**6.81**	**2.24**

Table 3: Experimental results of comparing M2OST with other ST or non-ST methods. The best results are marked in **bold**, and the second-best results are <u>underlined</u>.

replace CTMM, but it even fails to outperform the concatenation scheme.

Finally, we replaced CCMM with ordinary fully connected layers, and the FLOPs increased while the performance did not change much. This illustrates the effectiveness of CCMM in performing channel mixing for sequences of different lengths in M2OST.

Study on the Input Combinations for M2OST. Using M2OST as the backbone, various input combinations were fed into the model to verify the effectiveness of our many-to-one design. We kept the network width and depth identical for different combinations of inputs to ensure fairness during comparison, which also leads to similar parameter counts and FLOPs of the compared methods. The experimental results are summarized in Table 2.

Analysis of the table reveals that when employing M2OST as a one-to-one-based method, using level 0 pathology images yields optimal results across all three datasets. This is attributed to the comprehensive high-frequency information present in the level 0 pathology images, validating that the gene expression in a spot is primarily related to its corresponding tissue area. In this case, M2OST also did not surpass other one-to-one-based methods such as ResNet-50 and ST-Net when referring to the results in Table 3, which is mainly due to its smaller model size. Nonetheless, after introducing level 1 and level 2 image patches as additional inputs, the PCC of M2OST increases to 48.07%, 44.17%, and 50.50% on the three datasets, achieving state-of-the-art performance. This illustrates the effectiveness of the many-to-one scheme in M2OST, proving that introducing the multi-scale and surround-spot visual information for ST prediction can improve the model accuracy.

Experimental Results

Overview of the Experimental Results. The experimental results of the comparison between M2OST and other methods are presented in Table 3. This table provides detailed insights into the PCC and RMSE on various datasets of different methods, along with their parameter count and FLOPs. Analysis of the experimental results reveals that M2OST achieves superior performance with fewer FLOPs and a reduced parameter count. In comparison to ST-Net, which features 0.40M more parameters and 0.63G more FLOPs, M2OST surpasses its performance on HER2+ and cSCC datasets by 1.16% and 1.13% PCC, respectively.

Comparison between M2OST and One-to-one Multi-Scale Methods. In Table 3, we also have some comparisons with ordinary one-to-one multi-scale methods, such as CrossViT and HIPT/iStar. Compared with the vanilla ViT, CrossViT significant improvement in ST regression performance, confirming the value of incorporating multi-scale information for this task. However, since CrossViT is limited in its ability to fully utilize inter-spot information, it falls short of surpassing the performance of our proposed M2OST model.

In the case of iStar, the model achieved an even higher prediction accuracy for ST, underscoring the effectiveness of HIPT in extracting multi-scale features from WSIs. However, due to HIPT's hierarchical ViT architecture, training the model end-to-end is computationally expensive. As a result, iStar employs frozen HIPT weights to generate WSI features for ST prediction, which might compromise feature extraction performance. Furthermore, our observations (based on the official code release) indicate that iStar requires significantly more processing time during inference. This increased time is primarily attributed to its multi-scale feature extraction process, which operates patch by patch and scale by scale. When we limited M2OST's batch size to match iStar's GPU memory consumption, M2OST demonstrated an inference speed that was 100$\times$ faster than iStar's for ST regression. Despite this remarkable efficiency, M2OST still outperformed iStar, highlighting the superiority of end-to-end training in ST prediction and validating the effectiveness of our model design.

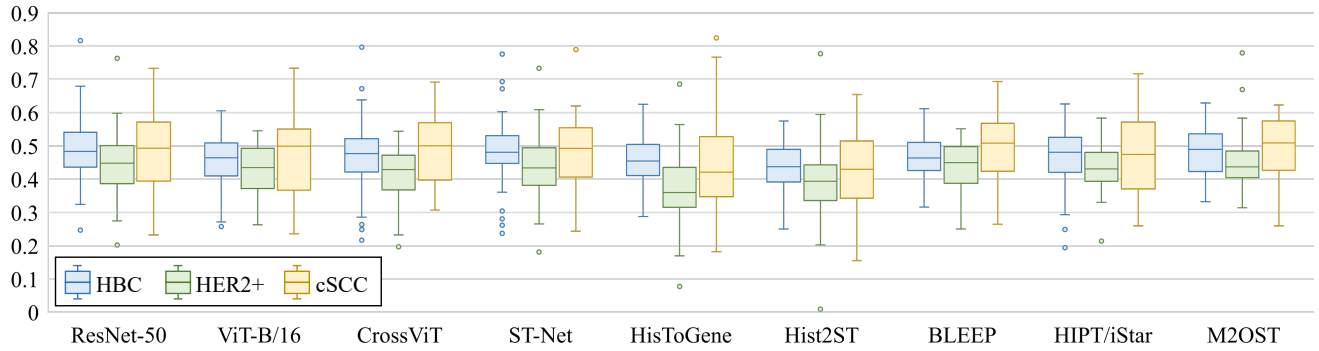

Figure 6: The box-plot of different methods' test PCC on the three datasets.

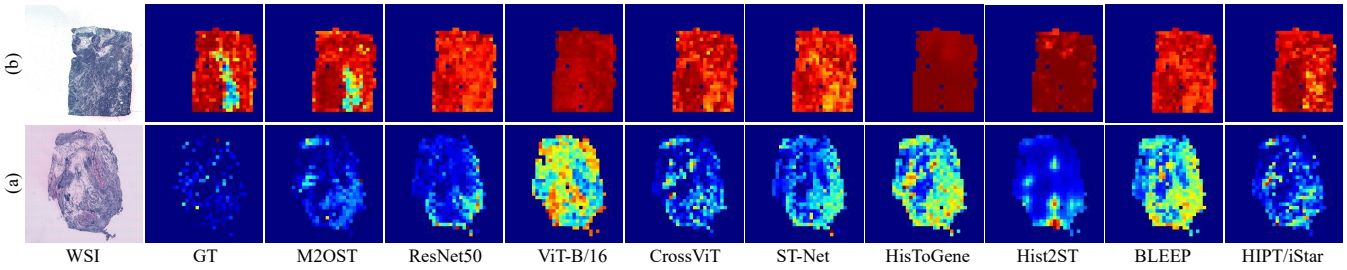

Figure 7: (a) Visualization of the ST map after PCA. (b) Visualization of the spatial distribution of the DDX5 gene.

Comparison between Patch-Level and Slide-Level ST Methods. From Table 3, it is also observed that the slide-level ST methods fail to outperform patch-level methods on all three datasets. Among the slide-level methods, Hist2ST does surpass HisToGene due to its larger model size, but the extra FLOPs and the dramatic parameter count diminish the significance of this performance improvement. When compared to baseline patch-level methods such as ST-Net, the PCC of Hist2ST is 2.78%, 2.99%, and 2.66% lower on the three datasets respectively. This suggests that the gene expressions of a spot are primarily related to its corresponding tissue area, and introducing inter-spot correlations does little to enhance prediction accuracy. Nevertheless, slide-level methods still possess the advantage of being more efficient in generating entire ST maps. With a refined network design, they still have the potential of achieving a competitive regression accuracy.

Statistical Significance and Deviation Analysis. A paired T-test for M2OST predictions has been conducted to ensure the statistical significance of the experimental results, and it is observed that p-value$<$0.05 holds for all other methods. We have also presented a boxplot in Fig 6, and it is shown that M2OST demonstrated the most stable predictions across all considered methods, validating the effectiveness of its network design.

Visualization Analysis. Finally, we present some visualization results in Figure 7 to make an intuitive comparison of the methods. In Figure 7(a), Principal Component Analysis (PCA) is used to compress the 250-dimension gene expressions into one dimension for better color mapping and visualization. As it is shown, slide-level methods such as HisToGene and Hist2ST tend to generate smoother ST maps, owing to the holistic processing of entire slides. In contrast, patch-level methods typically yield sharper predictions due to the independent processing of each spot in the ST map. Notably, M2OST consistently produces more accurate ST maps with distributions closely resembling the ground truth. This observation underscores the effectiveness of M2OST. Additionally, we augment our findings with individual gene visualizations in Figure 7(b) to further elucidate the efficacy of M2OST. The gene we selected for visualization is DDX5, which plays a pivotal role in the proliferation and tumorigenesis of non-small-cell cancer cells by activating the beta-catenin signaling pathway (Wang et al. 2015). Our results indicate that M2OST achieves the highest accuracy in gene expression prediction for the selected gene, surpassing the performance of other patch-level and slide-level methods.

Conclusion

In this study, we tackle the challenging task of predicting ST gene expressions from WSIs by proposing a novel many-to-one-based regression Transformer, M2OST. M2OST leverages pathology images from several distinct levels to collectively predict gene expressions within their common central tissue area. The model incorporates M2OST Encoder for decoupled multi-scale feature extraction, which comprises ITMM for intra-scale representation learning, CTMM for cross-scale feature extraction, and CCMM for multi-scale channel mixing. The experimental results on three public ST datasets show that M2OST can achieve state-of-the-art performance with minimal parameters and FLOPs.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (No. 2022YFC2504605). It was also supported in part by the Grant in Aid for Scientific Research from the Japanese Ministry for Education, Science, Culture and Sports (MEXT) under the Grant No. 20KK0234, 21H03470.

References

Andersson, A.; Larsson, L.; Stenbeck, L.; Salmén, F.; Ehinger, A.; Wu, S. Z.; Al-Eryani, G.; Roden, D.; Swarbrick, A.; Borg, Å.; et al. 2021. Spatial deconvolution of HER2-positive breast cancer delineates tumor-associated cell type interactions. *Nature communications*, 12(1): 6012.

Bressan, D.; Battistoni, G.; and Hannon, G. J. 2023. The dawn of spatial omics. *Science*, 381(6657): eabq4964.

Cang, Z.; Zhao, Y.; Almet, A. A.; Stabell, A.; Ramos, R.; Plikus, M. V.; Atwood, S. X.; and Nie, Q. 2023. Screening cell–cell communication in spatial transcriptomics via collective optimal transport. *Nature Methods*, 20(2): 218–228.

Chen, C.-F. R.; Fan, Q.; and Panda, R. 2021. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, 357–366.

Chen, R. J.; Chen, C.; Li, Y.; Chen, T. Y.; Trister, A. D.; Krishnan, R. G.; and Mahmood, F. 2022. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16144–16155.

Chung, Y.; Ha, J. H.; Im, K. C.; and Lee, J. S. 2024. Accurate Spatial Gene Expression Prediction by integrating Multi-resolution features. *arXiv preprint arXiv:2403.07592*.

Ding, K.; Zhou, M.; Metaxas, D. N.; and Zhang, S. 2023. Pathology-and-genomics multimodal transformer for survival outcome prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 622–631. Springer.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; and Lu, H. 2019. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3146–3154.

Hamilton, W.; Ying, Z.; and Leskovec, J. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30.

He, B.; Bergenstråhle, L.; Stenbeck, L.; Abid, A.; Andersson, A.; Borg, Å.; Maaskola, J.; Lundeberg, J.; and Zou, J. 2020. Integrating spatial gene expression and breast tumour morphology via deep learning. *Nature biomedical engineering*, 4(8): 827–834.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Hoang, D.-T.; Dinstag, G.; Hermida, L. C.; Ben-Zvi, D. S.; Elis, E.; Caley, K.; Sammut, S.-J.; Sinha, S.; Sinha, N.; Dampier, C. H.; et al. 2022. Prediction of cancer treatment response from histopathology images through imputed transcriptomics. *Research Square*.

Hu, C.; Xia, T.; Ju, S.; and Li, X. 2023. When sam meets medical images: An investigation of segment anything model (sam) on multi-phase liver tumor segmentation. *arXiv preprint arXiv:2304.08506*.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708.

Ji, A. L.; Rubin, A. J.; Thrane, K.; Jiang, S.; Reynolds, D. L.; Meyers, R. M.; Guo, M. G.; George, B. M.; Mollbrink, A.; Bergenstråhle, J.; et al. 2020. Multimodal analysis of composition and spatial architecture in human squamous cell carcinoma. *Cell*, 182(2): 497–514.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*.

Kolodziejczyk, A. A.; Kim, J. K.; Svensson, V.; Marioni, J. C.; and Teichmann, S. A. 2015. The technology and biology of single-cell RNA sequencing. *Molecular cell*, 58(4): 610–620.

Lee, Y.; Bogdanoff, D.; Wang, Y.; Hartoularos, G. C.; Woo, J. M.; Mowery, C. T.; Nisonoff, H. M.; Lee, D. S.; Sun, Y.; Lee, J.; et al. 2021. XYZeq: Spatially resolved single-cell RNA sequencing reveals expression heterogeneity in the tumor microenvironment. *Science advances*, 7(17): eabg4755.

Levy-Jurgenson, A.; Tekpli, X.; Kristensen, V. N.; and Yakhini, Z. 2020. Spatial transcriptomics inferred from pathology whole-slide images links tumor heterogeneity to survival in breast and lung cancer. *Scientific reports*, 10(1): 18802.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; and Xie, S. 2022. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11976–11986.

Monjo, T.; Koido, M.; Nagasawa, S.; Suzuki, Y.; and Kamatani, Y. 2022. Efficient prediction of a spatial transcriptomics profile better characterizes breast cancer tissue sections without costly experimentation. *Scientific Reports*, 12(1): 4133.

Mrabah, N.; Amar, M. M.; Bouguessa, M.; and Diallo, A. B. 2023. Toward convex manifolds: a geometric perspective for deep graph clustering of single-cell RNA-seq data. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 4855–4863.

Niazi, M. K. K.; Parwani, A. V.; and Gurcan, M. N. 2019. Digital pathology and artificial intelligence. *The lancet oncology*, 20(5): e253–e261.

Pang, M.; Su, K.; and Li, M. 2021. Leveraging information in spatial transcriptomics to predict super-resolution gene expression from histology images in tumors. *bioRxiv*, 2021–11.

Rao, A.; Barkley, D.; França, G. S.; and Yanai, I. 2021. Exploring tissue architecture using spatial transcriptomics. *Nature*, 596(7871): 211–220.

Rodriques, S. G.; Stickels, R. R.; Goeva, A.; Martin, C. A.; Murray, E.; Vanderburg, C. R.; Welch, J.; Chen, L. M.; Chen, F.; and Macosko, E. Z. 2019. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, 363(6434): 1463–1467.

Ryu, J.; Puche, A. V.; Shin, J.; Park, S.; Brattoli, B.; Lee, J.; Jung, W.; Cho, S. I.; Paeng, K.; Ock, C.-Y.; et al. 2023. OCELOT: Overlapped Cell on Tissue Dataset for Histopathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23902–23912.

Simonyan, K.; and Zisserman, A. 2015. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations (ICLR 2015)*. Computational and Biological Learning Society.

Song, Q.; and Su, J. 2021. DSTG: deconvoluting spatial transcriptomics data through graph-based artificial intelligence. *Briefings in bioinformatics*, 22(5): bbaa414.

Stenbeck, L.; Bergenstråhle, L.; Lundeberg, J.; and Borg, Å. 2021. Human breast cancer in situ capturing transcriptomics. *Mendeley Data*, 2.

Tian, L.; Chen, F.; and Macosko, E. Z. 2023. The expanding vistas of spatial transcriptomics. *Nature Biotechnology*, 41(6): 773–782.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, Z.; Luo, Z.; Zhou, L.; Li, X.; Jiang, T.; and Fu, E. 2015. DDX 5 promotes proliferation and tumorigenesis of non-small-cell lung cancer cells by activating β-catenin signaling pathway. *Cancer science*, 106(10): 1303–1312.

Wei, K.; Yang, Y.; Jin, L.; Sun, X.; Zhang, Z.; Zhang, J.; Li, X.; Zhang, L.; Liu, J.; and Zhi, G. 2023. Guide the Many-to-One Assignment: Open Information Extraction via IoU-aware Optimal Transport. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4971–4984.

Weitz, P.; Wang, Y.; Hartman, J.; and Rantalainen, M. 2021. An investigation of attention mechanisms in histopathology whole-slide-image analysis for regression objectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 611–619.

Xie, R.; Pang, K.; Chung, S.; Perciani, C.; MacParland, S.; Wang, B.; and Bader, G. 2024. Spatially Resolved Gene Expression Prediction from Histology Images via Bi-modal Contrastive Learning. *Advances in Neural Information Processing Systems*, 36.

Yarlagadda, D. V. K.; Massagué, J.; and Leslie, C. 2023. Discrete Representation Learning for Modeling Imaging-based Spatial Transcriptomics Data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3846–3855.

Zehui, L.; Liu, P.; Huang, L.; Chen, J.; Qiu, X.; and Huang, X. 2019. Dropattention: A regularization method for fully-connected self-attention networks. *arXiv preprint arXiv:1907.11065*.

Zeng, Y.; Wei, Z.; Yu, W.; Yin, R.; Yuan, Y.; Li, B.; Tang, Z.; Lu, Y.; and Yang, Y. 2022. Spatial transcriptomics prediction from histology jointly through transformer and graph neural networks. *Briefings in Bioinformatics*, 23(5): bbac297.

Zhang, D.; Schroeder, A.; Yan, H.; Yang, H.; Hu, J.; Lee, M. Y.; Cho, K. S.; Susztak, K.; Xu, G. X.; Feldman, M. D.; et al. 2024. Inferring super-resolution tissue architecture by integrating spatial transcriptomics with histology. *Nature Biotechnology*, 1–6.