Medical Image Analysis 108 (2026) 103889

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Spatial transcriptomics expression prediction from histopathology based on cross-modal mask reconstruction and contrastive learning

Junzhuo Liu [a], Markus Eckstein [b,c,d,e], Zhixiang Wang [f], Friedrich Feuerhake [g,h], Dorit Merhof [a,i,*]

[a] Faculty of Informatics and Data Science, University of Regensburg, Regensburg, Germany
[b] Institute of Pathology, Universitätsklinikum Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany
[c] Comprehensive Cancer Center Erlangen-EMN (CCC ER-EMN), Erlangen, Germany
[d] Comprehensive Cancer Center Alliance WERA (CCC WERA), Erlangen, Germany
[e] Bavarian Cancer Research Center (BZKF), Erlangen, Germany
[f] Department of Ultrasound, Beijing Friendship Hospital, Capital Medical University, Beijing, China
[g] Institute for Pathology, Hanover Medical School, Hannover, Germany
[h] Institute for Neuropathology, University Clinic Freiburg, Freiburg im Breisgau, Germany
[i] Fraunhofer Institute for Digital Medicine MEVIS, Bremen, Germany

ARTICLE INFO

Keywords:
Histopathology
Spatial transcriptomics
Contrastive learning
Multimodal fusion

ABSTRACT

Spatial transcriptomics is a technology that captures gene expression at different spatial locations, widely used in tumor microenvironment analysis and molecular profiling of histopathology, providing valuable insights into resolving gene expression and clinical diagnosis of cancer. Due to the high cost of data acquisition, large-scale spatial transcriptomics data remain challenging to obtain. In this study, we develop a contrastive learning-based deep learning method to predict spatially resolved gene expression from the whole-slide images (WSIs). Unlike existing end-to-end prediction frameworks, our method leverages multi-modal contrastive learning to establish a correspondence between histopathological morphology and spatial gene expression in the feature space. By computing cross-modal feature similarity, our method generates spatially resolved gene expression directly from WSIs. Furthermore, to enhance the standard contrastive learning paradigm, a cross-modal masked reconstruction is designed as a pretext task, enabling feature-level fusion between modalities. Notably, our method does not rely on large-scale pretraining datasets or abstract semantic representations from either modality, making it particularly effective for scenarios with limited spatial transcriptomics data. Evaluation across six different disease datasets demonstrates that, compared to existing studies, our method improves Pearson Correlation Coefficient (PCC) in the prediction of highly expressed genes, highly variable genes, and marker genes by 6.27 %, 6.11 %, and 11.26 % respectively. Further analysis indicates that our method preserves gene-gene correlations and applies to datasets with limited samples. Additionally, our method exhibits potential in cancer tissue localization based on biomarker expression. The code repository for this work is available at https://github.com/ngfufdrdh/CMRCNet.

1. Introduction

Cancer is one of the leading causes of mortality worldwide, with approximately one in five men or women developing cancer during their lifetime (Bray et al., 2024). In 2020, nearly 20 million new cancer cases and approximately 10 million cancer-related deaths were reported globally. Demographic projections indicate a rising trend in these numbers, continuously imposing significant burdens on public health and socioeconomic systems (Sung et al., 2021). In clinical practice, cancer is usually diagnosed based on microscopic analysis of tissue obtained by surgical resection or biopsy. The current standard for tumor classification defined by the WHO is based on histopathology and increasingly on additional molecular markers such as DNA sequencing, fluorescence in-situ hybridization (FISH), and DNA methylation analysis, defining histogenesis (cellular origin) and degree of malignancy (histopathological grade) (Kleihues et al., 2002). In addition, the tumor

* Corresponding author.
E-mail address: Dorit.Merhof@informatik.uni-regensburg.de (D. Merhof).

https://doi.org/10.1016/j.media.2025.103889
Received 16 June 2025; Received in revised form 23 October 2025; Accepted 24 November 2025
Available online 25 November 2025

size and its spreading into surrounding tissues (locally invasive growth) and distant organs (metastatic spread) are classified by the IUCC in the TNM classification system (Brierley et al., 2017). Cancer is a complex disease (Fu et al., 2020) caused by dysregulation of multiple cellular functions referred to as "hallmarks of cancer" (Swanton et al., 2024). In specific types of cancer, histopathology provides an intuitive visualization of tumor cell morphology (Ektefaie et al., 2021; Hölscher et al., 2023; Sparks and Madabhushi, 2013), tissue architecture (Hu et al., 2021) and differentiation status (Rakha et al., 2022; Takahara et al., 2021), offering critical clinical insights for cancer diagnosis (Kilim et al., 2025; Tsuneki et al., 2023), tumor grading (Handley et al., 2022) and prognostic assessment (Cannella et al., 2024; Ho et al., 2024). The advent of spatial transcriptomics (ST) has provided a molecular-level complement to histopathology-based diagnostics. ST not only reveals spatial transcriptional patterns and regulatory mechanisms within tissues (Garcia-Alonso et al., 2021), but also captures the tumor microenvironment and local tissue characteristics (Chen et al., 2020c; Williams et al., 2022), offering novel insights for precision oncology.

The acquisition of spatial transcriptomics data typically involves tissue sample sectioning, tissue dissociation, and the integration of barcodes and indices (Jin et al., 2024). The high cost of specialized equipment and the complexity of the preparation process pose significant challenges for generating large-scale spatial transcriptomic data from WSIs in clinical practice (Yu et al., 2022). Therefore, deep learning models that predict spatial transcriptomic expression directly from histopathology images offer a promising solution to this limitation. STNet (He et al., 2020a) was one of the pioneering methods for modeling the correlation between tissue morphology and spatial transcriptomic expression using convolutional neural networks. It employed a convolutional architecture to hierarchically extract deep visual features from histopathological images and mapped these features to spatial gene expression in an end-to-end manner. HisToGene (Pang et al., 2021) enhanced spatial dependency modeling of spatial transcriptomic measurement spots by leveraging a Transformer-based framework. Similarly, EGGN (Yang et al., 2024) utilized graph neural networks (GNNs) to model variations in gene expression levels across different spatial windows, dynamically improving gene expression prediction.

Despite the promising potential of these methods in predicting spatially resolved gene expression by transcriptomics from histopathological images, several limitations and challenges remain: 1. Most end-to-end models inevitably rely on fixed gene labels (Jaume et al., 2025; Rahaman et al., 2023). These methods select target genes based on their average expression levels or variability across the entire dataset. However, even within the same dataset, spatial transcriptomic gene expression levels and variability can vary significantly among different samples. As a result, genes that exhibit high expression or variability only in a subset of samples may not be included in the fixed prediction targets, thereby reducing the flexibility of gene expression prediction. Recently, BLEEP (Xie et al., 2024) was introduced to improve prediction flexibility based on contrastive learning. BLEEP models the relationship between histopathological morphological features and spatial transcriptomic expression through contrastive learning by computing the similarity between cross-modal data in the feature space to generate gene expression predictions from histopathological images. However, previous studies suggest that multimodal features often reside in distinct representational spaces after being encoded by multimodal encoders, making it challenging for the encoder to capture their interactions effectively (Li et al., 2021). Moreover, the inherent discreteness of spatial gene expression sequences and the absence of explicit semantic representations make it difficult to employ semantic-based strategies, such as cross-modal semantic matching, to capture multimodal interactions. 2. Current methods have not proposed effective strategies for integrating histopathological images with spatial transcriptomic data. Existing end-to-end models treat spatial transcriptomic data merely as labels while using histopathology images as input. BLEEP (Xie et al., 2024) implements cross-modal contrastive learning based on an individual discrimination task (Wu et al., 2018). However, it relies solely on the global similarity between cross-modal embeddings to optimize the feature space, without performing more fine-grained alignment and fusion. 3. Most existing approaches are validated on only a single disease dataset and do not adopt cross-validation (He et al., 2020a; Pang et al., 2021). Our research indicates that, even within the same dataset cohort, different samples exhibit varying degrees of prediction difficulty. Validation based on only a subset of samples often results in unreliable performance evaluation.

To address these limitations and to model the relationship between histopathological morphological features and spatial transcriptomic expression for gene expression prediction from images, we propose a novel contrastive learning framework based on cross-modal mask reconstruction, named CMRCNet. CMRCNet comprises three stages: contrastive learning, cross-modal reconstruction, and similarity-based inference. In the contrastive learning stage, an image encoder and a gene encoder extract features from paired histopathological image patches and spatial transcriptomic expression data, respectively, constructing embeddings for the similarity matrix used in contrastive learning. The contrastive loss function serves as an initial mechanism to bring the embeddings of the two modalities closer in the feature space. To achieve further alignment of the embeddings at local and fine-grained levels, we introduce cross-modal reconstruction as a pretext task. Unlike approaches that are explicitly semantics-driven, cross-modal reconstruction can be viewed as a set of local, pixel-level supervisory tasks that enforce the alignment of morphological and gene expression features in the feature space. Specifically, in the cross-modal reconstruction stage, randomly masked image features are progressively reconstructed under the guidance of gene expression features. We propose a cross-modal reconstruction module based on a Transformer architecture to ensure consistency in image feature encoding. Within this module, we design a cross-attention mechanism with residual mapping to restore image features from gene encodings. The proposed attention mechanism not only prevents the reconstruction process from overly relying on a single modality, but also ensures that unmasked morphological features and fused information at each stage contribute to the reconstruction. We evaluate our method on six datasets across different diseases, predicting highly expressed genes, highly variable genes, and marker genes, demonstrating the superiority of CMRCNet. CMRCNet not only achieves a higher similarity to the real spatial transcriptomic expression but also exhibits significant potential in preserving biological heterogeneity and predicting tumor regions based on marker genes.

The main contributions of this work are as follows:

1. We propose to use cross-modal reconstruction as a pretext task to enhance image-to-spatial transcriptomics contrastive learning, enabling the prediction of gene expression levels from histopathological features. Unlike most contrastive learning variants, our proposed CMRCNet does not rely on explicit semantic information or large-scale training datasets.
2. We propose a cross-modal fusion module to facilitate feature interactions between histomorphological representations and gene expression in contrastive learning. This module effectively balances the contribution of different modalities during fusion, preventing over-reliance on a single modality.
3. We conduct extensive experiments on six datasets covering different diseases using multi-fold cross-validation, mitigating the impact of varying prediction difficulty across samples on result reliability. Beyond predicting highly variable and highly expressed genes, we provide evaluations on marker gene predictions, further revealing the potential of leveraging artificial intelligence to integrate histopathology and spatial transcriptomics for clinical cancer diagnosis.

2. Related work

2.1. Spatial transcriptomics prediction

Spatial transcriptomics is a technique that captures gene expression levels at different spatial locations and is widely used in tumor microenvironment analysis (Hunter et al., 2021; Khaliq et al., 2024) and molecular profiling of histopathology (Moncada et al., 2020; Ståhl et al., 2016; Wang et al., 2023). Spatial transcriptomics alleviates the limitations of single-cell RNA sequencing (scRNA-seq), which requires the release of viable cells from the entire tissue without inducing stress, cell death, and/or cell aggregation (Williams et al., 2022), thereby providing valuable insights into spatially resolved gene expression analysis within tissues. The interaction between artificial intelligence and molecular data is an emerging field. For example, BioGPT (Al-Kateb et al., 2025) integrates gene expression with clinical reports to support classification and interpretable diagnosis. In response to the widespread challenge of limited accessibility to spatial gene expression data, current deep learning-based approaches for predicting spatial transcriptomic expression from histopathological images are categorized into two paradigms: end-to-end direct prediction and contrastive learning-based prediction.

The end-to-end paradigm formulates the prediction task as a multivariate regression (Mejia et al., 2023). It takes histopathology images as input and employs deep neural networks to map visual features to the expression levels of target genes. STNet (He et al., 2020a) is one of the most representative methods in this category. It segments high-resolution histopathology images into individual patches, extracts visual features from these patches using DenseNet-121 (Huang et al., 2017), and employs linear layers to map deep features to the predicted expression levels of 250 genes. Similar approaches include BrSTNet (Rahaman et al., 2023), EGN (Yang et al., 2023), EGGN (Yang et al., 2024), and SEPAL (Mejia et al., 2023). BrSTNet focuses on breast histopathology images and introduces an auxiliary network that operates in parallel with the main network. The linear layers in the main network predict the expression of 250 highly expressed genes, while those in the auxiliary network predict the remaining genes. Additionally, BrSTNet incorporates SSAL (Palacio et al., 2021) regularization to enhance the generalizability of the main network. EGN and EGGN both leverage similar patches within histopathology slides to dynamically enhance gene expression prediction. These methods are based on the assumption that patches with similar tissue morphology exhibit similar gene expression patterns. EGN employs a Vision Transformer (Dosovitskiy, 2020) to extract visual features from patches and utilizes a specialized instance retrieval mechanism to refine the intermediate representations of the main encoder by adaptively incorporating the nearest instances in the feature space. In contrast, EGGN first constructs a similarity graph across different spatial windows. This similarity graph is then processed by GraphSAGE (Hamilton et al., 2017) to extract features and facilitate information propagation between windows. However, this approach, which relies on visually similar patches, overlooks the local contextual information surrounding individual patches. The tissue background in which a patch is situated may influence gene expression levels. To address this limitation, SEPAL integrates local contextual information using a graph neural network, achieving superior gene expression prediction performance on breast cancer datasets.

Unlike the end-to-end paradigm, contrastive learning-based methods focus on establishing feature associations between histopathology patches and spatial transcriptomic expressions. These methods take both histopathology images and spatial transcriptomic data as input and employ instance discrimination (Wu et al., 2018) as a pretext task to construct contrastive loss functions. BLEEP (Xie et al., 2024) is the first method to leverage contrastive learning for predicting spatial transcriptomic expression from histopathology images. During the training phase, ResNet50 and a linear mapping module serve as the image and gene encoders, extracting features from tissue patches and transcriptomic spots. Feature sequences from different patches and transcriptomic spots are used to construct a similarity matrix via matrix multiplication. The contrastive loss is supervised by intra-modal similarity matrices derived from patch and transcriptomic spot features. The objective of BLEEP is to bring the visual representations of histopathology images and their corresponding gene representations closer together in the feature space. During inference, for each tissue patch in the test set, BLEEP selects the 50 most similar gene representations from the training set as keys based on feature similarity. The mean expression of these keys is used as the predicted gene expression for the corresponding patch. Similar to BLEEP, RankByGene (Huang et al., 2024) improves contrastive learning by leveraging the relative consistency of similarity rankings across different modalities and instances. Compared to end-to-end models, a significant advantage of these methods is their flexibility. They do not require predefined labels, but instead dynamically predict highly expressed or highly variable genes during inference based on different test samples. This adaptability is particularly important, as highly expressed or highly variable genes often differ across samples, even within the same disease cohort.

2.2. Contrastive learning

Contrastive learning is a feature learning approach based on sample similarity, grounded in the fundamental assumption that similar samples should be closer in the feature space. Typically, contrastive learning requires the construction of positive and negative samples for each instance. By learning the similarities and differences between positive and negative sample pairs, it encourages the encoder to capture meaningful data representations. Since contrastive learning does not inherently rely on fixed labels, it requires a pretext task to define the loss function. Instance discrimination (Wu et al., 2018) is one of the earliest contrastive learning methods. It treats each sample in the dataset as a unique class and trains an encoder to learn representations that distinguish each sample as much as possible. The positive samples are generated via random data augmentations of the same sample, while the negative samples are drawn from other samples in the dataset and stored in a memory bank. Similar approaches have been explored in (Caron et al., 2020; Chen et al., 2020b; He et al., 2020b). Unlike the positive sample construction in (Wu et al., 2018), CMC (Tian et al., 2020) proposes that different viewpoints of the same object in an image can be considered as a positive sample pair. The goal of CMC is to maximize the mutual information between different viewpoints, enabling the encoder to learn object-intrinsic features while minimizing interference from viewpoint variations and background noise. CLIP (Radford et al., 2021) is one of the earliest studies that applied contrastive learning to multimodal settings. It extends the positive-negative sample pair design from intra-modal to cross-modal learning, particularly between images and text. An image and its corresponding textual description are treated as a positive sample pair. Leveraging large-scale training data and multimodal encoders with similar architectures, CLIP has achieved remarkable success in various domains, including image retrieval (Baldrati et al., 2022; Luo et al., 2022; Ma et al., 2022) and text-to-image generation (Tao et al., 2023; Wang et al., 2022b).

3. Method

3.1. CMRCNet architecture

Our novel method (CMRCNet) is inspired by CLIP (Radford et al., 2021) and BLEEP (Xie et al., 2024). However, CLIP is primarily designed for large-scale contrastive learning between images and text, and BLEEP focuses solely on narrowing the spatial gap between pathological morphology and spatial gene expression, without enabling deeper cross-modal feature interactions. In contrast, CMRCNet utilizes cross-modal reconstruction to achieve interaction and alignment of histopathological morphology and gene expression levels at the feature

pixel level, and is suitable for small-scale datasets. The overall workflow of CMRCNet is illustrated in Fig. 1. It consists of three main stages: contrastive learning, cross-modal reconstruction, and similarity-based retrieval during inference. During the training phase, the model takes paired histopathology image patches and spatial transcriptomic expression profiles as input. In the inference phase, histopathology images from the test set are fed into the image encoder, and spatial transcriptomic data from the training set are used as input to the gene encoder. The final spatial transcriptomic expression predictions are obtained through similarity-based retrieval.

3.2. Contrastive learning

Each whole slide image $X \in \mathbb{R}^{H\times W\times 3}$ is first divided into patches $I \in \mathbb{R}^{N\times L\times L\times 3}$, where N represents the number of patches, and L denotes the patch resolution. In contrastive learning, the inputs for the image encoder and transcriptomic encoder are $I \in \mathbb{R}^{B\times L\times L\times 3}$ and $S \in \mathbb{R}^{B\times d}$, where B represents the batch size, set to 64 in our method, and d denotes the dimension of gene expression in spatial transcriptomics. To ensure consistency between the architectures of the image encoder in contrastive learning and the cross-attention module in cross-modal reconstruction, we employ the Vision Transformer (ViT) base version as the image encoder. For a single image patch $I \in \mathbb{R}^{1\times L\times L\times 3}$ with $L = 224$, ViT first partitions it into non-overlapping smaller patches of size 32×32, resulting in a total of 49 patches. Each patch is then flattened and linearly projected into a 768-dimensional embedding vector (token). To encode spatial positional information, a positional token is added. The resulting sequence of 50 tokens is fed into the ViT. The ViT is composed of multiple Transformer blocks, each consisting of a multi-head self-attention mechanism and a feed-forward network. The multi-head self-attention mechanism facilitates information exchange across patches, enabling effective global modeling. The feed-forward network further

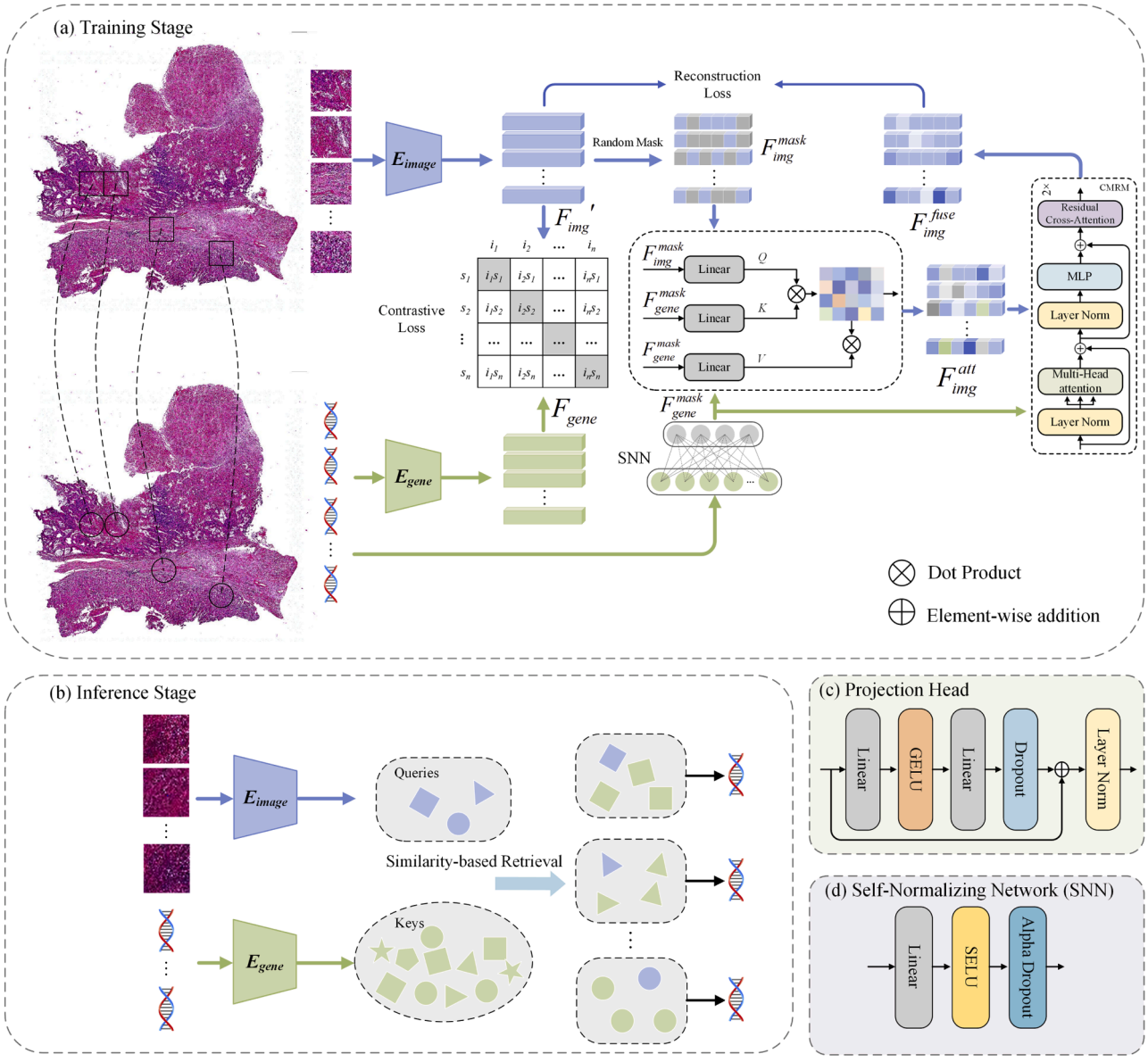

Fig. 1. Overview of CMRCNet. (a) Training Stage: The input data for CMRCNet in the training stage are histopathological patches and transcript gene expression, including contrastive loss and reconstruction loss. (b) Inference Stage: The input data for CMRCNet in the inference stage are histopathological patches from the test set and transcript gene expression from the train set, and the prediction results are obtained through similarity-based retrieval. (c) The Projection Head is used to map high-dimensional features to lower dimensions for contrastive learning. (d) The self-normalizing network is used to map spatial transcriptomics data to lower-dimensional features for constructing reconstruction loss.

processes the features from the attention layer, enhancing their representational capacity through nonlinear transformations. After feature extraction, the output of the image encoder is $F_{img} \in \mathbb{R}^{B\times\ 50\ \times\ 768}$, where 50 represents the number of tokens, and 768 is the embedding dimension of ViT-B. Next, the average value of F_{img} is computed along the token dimension and fed into the projection head, which maps the high-dimensional hidden representations from the encoder to a lower-dimensional space. The projection head consists of two fully connected layers, a GELU activation function, dropout, and layer normalization, which is consistent with the design in BLEEP. The architecture of the projection head is shown in Fig. 1(c). The output of the projection head is $F'_{img} \in \mathbb{R}^{B\times\ 256}$, which serves as the feature representation for contrastive learning. For spatial transcriptomic expression $S \in \mathbb{R}^{B\times d}$, the projection head serves as the gene encoder, extracting features from the expression data, with the output given by $F_{gene} \in \mathbb{R}^{B\times\ 256}$.

We follow the construction of the contrastive loss as implemented in BLEEP (Xie et al., 2024), where similarity computation and label construction are performed within a mini-batch. For the image modality and the gene modality, we first compute the intra-modal similarities separately to capture the similarity relationships within each modality. For image features $F'_{img} \in \mathbb{R}^{B\times\ 256}$ and gene expression features $F_{gene} \in \mathbb{R}^{B\times\ 256}$, the intra-modal similarity matrices are computed as:

$$sim\left(F'_{img},\ F'_{img}\right) \in \mathbb{R}^{B\times B} = F'_{img} \cdot {F'_{img}}^{T} \tag{1}$$

$$sim\left(F_{gene},\ F_{gene}\right) \in \mathbb{R}^{B\times B} = F_{gene} \cdot F_{gene}^{T} \tag{2}$$

Here, each entry in the similarity matrix represents the pairwise similarity score between two samples within the same modality. Next, we compute the similarity between the two modalities, which explicitly models the alignment between paired image and gene expression representations. This is achieved by computing the dot-product similarities in both directions, as shown in the following equation:

$$sim\left(F'_{img},\ F_{gene}\right) \in \mathbb{R}^{B\times B} = F'_{img} \cdot F_{gene}^{T} \tag{3}$$

$$sim\left(F_{gene},\ F'_{img}\right) \in \mathbb{R}^{B\times B} = F_{gene} \cdot {F'_{img}}^{T} \tag{4}$$

Each row of these similarity matrices corresponds to the similarity distribution of one sample with respect to all samples from the other modality. The label construction process for contrastive learning is shown in the following equation:

$$target = softmax\left(\frac{sim\left(F'_{img},\ F'_{img}\right) + sim\left(F_{gene},\ F_{gene}\right)}{2} \times \tau\right) \tag{5}$$

τ is the temperature function in contrastive learning. We follow the settings from BLEEP, where $\tau = 1.0$. Cross-entropy (CE) is used as the contrastive learning loss function, modeling the correlation between the two modalities based on similarity calculation. The overall contrastive learning loss function is defined as follows:

$$Loss_c = CE\left(sim\left(F'_{img},\ F_{gene}\right),\ target\right) + CE\left(sim\left(F_{gene},\ F'_{img}\right),\ target^{T}\right) \tag{6}$$

The first term aligns the image-to-gene similarities with the target distribution, while the second term aligns the gene-to-image similarities. Together, they enforce bidirectional consistency across modalities, ensuring that paired image and gene expression features are embedded close to each other in the shared feature space.

3.3. Cross-modal reconstruction

The objective of multimodal contrastive learning is to minimize the distance between paired or similar multimodal samples in the feature space while keeping unrelated samples apart. It relies on a loss function defined by feature similarity, which encourages two similar embeddings to be as close as possible in the feature space. However, this loss primarily considers the global similarity between two embeddings, without accounting for their internal structures or enforcing a more fine-grained alignment between them. Thanks to the rich semantic information and large training datasets, improved versions of CLIP propose leveraging cross-modal semantic matching or hard negative mining to improve contrastive learning. Compared with text, spatial transcriptomics data exhibit highly discrete gene sequences and lack explicit semantic information, making architectures based on cross-modal semantic matching or hard negative mining difficult to transfer to morphology–spatial gene expression contrastive learning. To enable the encoder to learn cross-modal interactions further and promote the alignment of the two modalities, inspired by MAE (He et al., 2022), we propose a cross-modal reconstruction auxiliary task to achieve feature-level fusion between tissue morphological features and gene expression levels. Unlike MAE, which trains a robust image encoder through unimodal self-reconstruction of randomly masked image patches, our cross-modal reconstruction strategy operates across multiple modalities and hidden-layer embeddings. This enhances the interaction and alignment of multimodal features.

The motivations of cross-modal reconstruction are as follows: (1) It provides denser and more localized supervision signals, which facilitate the alignment of multimodal embeddings. Masked reconstruction requires the model to perform pixel-wise predictions in the masked regions. This compels embeddings from different modalities to be aligned at a finer granularity. (2) Cross-modal reconstruction is particularly suitable for morphology–spatial gene expression contrastive learning. Cross-modal reconstruction encourages the model to learn fine-grained mappings between the two modalities, thereby promoting deeper alignment and fusion. This alignment is not dependent on explicit semantics but is instead feature-based. Compared with contrastive losses that only consider global similarity, cross-modal reconstruction emphasizes local correspondences and inter-modality associations, which is beneficial for uncovering the complex relationships between morphology and gene expression. Furthermore, it can be viewed as multiple local prediction tasks with fixed targets for each sample, which increases the amount and stability of effective supervision, thereby benefiting contrastive learning on small-scale spatial transcriptomics data.

For the output of the image encoder $F_{img} \in \mathbb{R}^{B\times\ 50\ \times\ 768}$, we use a fully connected layer to map the embedding dimension from 768 to 256. For the discrete gene data, we do not employ a Transformer encoder. On one hand, the contextual and global information in the original gene sequences is very limited. On the other hand, it introduces additional parameters and increases the risk of overfitting. We utilize Self-Normalizing Networks (SNN) (Klambauer et al., 2017) to map the transcript expression from the initial dimension to 2048 dimensions, and then perform a dimensionality transformation to construct the transcriptomic features $F_{gene}^{mask} \in \mathbb{R}^{B\times\ 8\ \times\ 256}$. The SNN, which is designed to encode gene expression sequences (Chen et al., 2020a; Ding et al., 2023), consists of fully connected layers, a SELU activation function, and Alpha Dropout with a dropout probability of 0.25. The architecture of SNN is shown in Fig. 1(d). The SELU is continuous and differentiable at all points, avoiding the issue of neuron death, and causes the output to tend towards zero mean and unit variance. Combined with SELU, Alpha Dropout further maintains the self-normalizing property of the output, which alleviates overfitting for high-dimensional but few-shot gene data. First, we randomly mask 50 % of the feature pixels in the image features, and the masked feature representation is denoted as $F_{img}^{mask} \in \mathbb{R}^{B\times\ 50\ \times\ 256}$. In our ablation evaluation, the combination of a 50 % mask rate and MSE reconstruction loss function shows the best performance. Next, the masked image features and transcriptomic features are input into the cross-attention module for preliminary reconstruction.

Since image features serve as the main feature for reconstruction, in the preliminary reconstruction process, the Query (Q) is generated from F_{img}^{mask}, while the Key (K) and Value (V) are generated from F_{gene}^{mask} The construction process of Q, K, and V is as follows:

$$Q = F_{img}^{mask} \cdot W_Q \tag{7}$$

$$K = F_{gene}^{mask} \cdot W_K \tag{8}$$

$$V = F_{gene}^{mask} \cdot W_V \tag{9}$$

The output of the cross-attention is denoted as $F_{img}^{att} \in \mathbb{R}^{B\times\ 50\ \times\ 256}$. This output is calculated as

$$F_{img}^{att} = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{10}$$

where d_k represents the dimension of the feature embeddings. Cross-attention is implemented using the multi-head approach, with the number of heads set to 4. Next, F_{img}^{att} is fed into two concatenated cross-modal fusion modules for feature extraction and further cross-modal reconstruction. Our cross-modal fusion module consists of two steps: image feature extraction and the fusion of image and gene expression features. To maintain consistency with the image encoder feature extraction, the cross-modal fusion module adopts the structure of Transformer blocks, which includes layer normalization, the multi-head self-attention mechanism, and linear mapping layers. In the cross-modal fusion module, the number of heads is set to 8. A cross-attention module is added as the final layer of the cross-modal fusion module. The architecture of the cross-attention module is shown in Fig. 2. Unlike cross-attention in the initial reconstruction, we also generate Value (V) from the image features and add it to the features weighted by the attention matrix in a residual-connected manner. This design has two motivations: (1) It prevents the information from overly favoring the spatial transcriptomic modality. Spatial transcriptomic data inevitably contain noise. When features overly rely on gene modality representations in the reconstruction process, noise is often retained in the fused features, making optimization difficult. (2) It preserves feature representations from the initial fusion and self-reconstruction. Before entering the cross-attention mechanism, image features undergo initial reconstruction and complete their information interaction within the Transformer block. The mappings from these features are added to the final feature using a residual connection. This ensures that the reconstruction is not solely dependent on the gene expression features. Unmasked tissue morphology features and fused features from each stage also contribute to the reconstruction process. The output of the cross-modal fusion module is denoted as $F_{fuse} \in \mathbb{R}^{B\times\ 50\ \times\ 256}$.

In the cross-modal reconstruction task, we use the MSE loss function to guide the model's learning. The image feature embeddings used for contrastive learning serve as the reconstruction target. F_{fuse} is averaged along the token dimension to produce the reconstruction result, denoted as $F_r \in \mathbb{R}^{B\times\ 256}$. The overall loss function for the cross-modal reconstruction is as follows:

$$Loss_r = MSE\left(F_r, F_{img}^{'}\right) \tag{11}$$

3.4. Loss function

Our method includes two loss functions: the contrastive learning loss and the cross-modal reconstruction loss. Since the two loss functions differ in magnitude, we use a logarithmic function to balance the optimization process and prevent one loss from dominating the model training. The overall loss function is as follows:

$$Loss_{total} = \log(1 + Loss_c) + \log(1 + Loss_r) \tag{12}$$

3.5. Similarity-based retrieval in the inference stage

The inference of spatial transcriptomic expression from images is completed based on similarity retrieval. The process of inference is shown in Fig. 1(b). During the inference stage, we only retain the VIT and projection head used for image encoding in contrastive learning, as well as the projection head used for spatial transcriptomic encoding. First, the WSIs from the test dataset are divided into patches based on the spatial transcriptomic location information and input into the VIT and Projection Head, producing image feature embeddings $F_{img}^{test} \in \mathbb{R}^{B\times\ 256}$. Next, the spatial transcriptomic data from the training dataset is used as input for the trained gene encoder, producing gene feature embeddings $F_{gene}^{test} \in \mathbb{R}^{M\times\ 256}$, where M represents the total number of transcriptomic spots in the training dataset. Then, the image feature embeddings and gene feature embeddings are used to compute the cosine similarity, resulting in a similarity matrix $S \in \mathbb{R}^{B\times M}$. The top 50 gene feature embeddings most similar to the current image feature embedding are selected as keys. The average gene expression values of the corresponding transcriptomic spots are taken as the final prediction results.

4. Experiments and results

4.1. Datasets

To validate the effectiveness of the proposed method, we train and evaluate CMRCNet on six datasets reflecting different diseases, including axillary lymph nodes in invasive ductal carcinoma (IDC-LymphNode) (Liu et al., 2022), skin cutaneous melanoma (SKCM) (Jaume et al., 2025), rectal adenocarcinoma (READ) (Valdeolivas et al., 2023), colonic adenocarcinoma (COAD) (Oliveira et al., 2024), invasive ductal carcinoma (IDC) (Janesick et al., 2023), and primary sclerosing cholangitis (PSC) (Andrews et al., 2024). All these datasets are collected

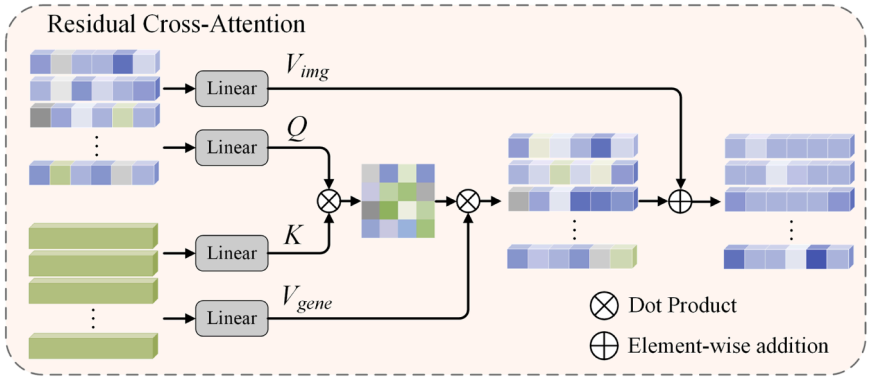

Fig. 2. Overview of residual cross-attention module. It is used to fuse histopathological features and gene expression features in the cross-modal reconstruction stage.

and curated by the HEST-benchmark (Jaume et al., 2025): https://huggingface.co/datasets/MahmoodLab/hest.

IDC-LymphNode (Liu et al., 2022): Lymph node–positive invasive ductal carcinoma. This dataset consists of four whole-slide images (WSIs) and spatial transcriptomics samples from patients diagnosed with breast cancer. The spatial transcriptomics data are obtained using the Visium transcriptomics technology (Ståhl et al., 2016), containing a total of 19,964 transcriptomic spots and expression levels of 33,921 genes. Scanpy (Wolf et al., 2018) is used to select the top 1000 highly variable genes from each sample, resulting in a final dataset comprising 3467 genes. For the evaluation of marker genes and genes influencing disease progression, the following genes are included in our study: ZNF276, ZNF750, YBX1, VHL, RUNX2, RND, ERCC4, PADI2, S100A8, and S100A6.

SKCM (Jaume et al., 2025): Skin cutaneous melanoma. This dataset includes two samples from two patients with skin cancer. The spatial transcriptomic data are obtained using Xenium technology, which contains 5716 transcriptomic spots and the expression levels of 541 genes. All of these genes are included in our study. We select MIA, PMEL, MLANA, MITF, and TYR as the marker genes for the SKCM dataset.

READ (Valdeolivas et al., 2023): Rectum adenocarcinoma. This dataset consists of four samples from two patients with rectal adenocarcinoma. The spatial transcriptomic data are obtained using Visium, containing 8407 transcriptomic spots and expression levels of 36,601 genes. Similar to the preprocessing steps of the IDC-LymphNodes dataset, we select 3355 highly variable genes. The marker genes for the READ dataset include: OLFM4, PLA2G2A, PLA2G2D, MMP1, MMP11, MMP12, SMAD4, MUC13, SLC26A3, ZG16 and NDRG1.

COAD (Oliveira et al., 2024): Colon adenocarcinoma. This dataset includes four WSIs and spatial transcript samples from patients with colonic adenocarcinoma. The spatial transcriptomic data are obtained using Xenium technology, comprising 18,523 transcriptomic spots and the expression levels of 541 genes. After excluding genes with different names and blank-named genes, 412 genes are included in our study. The marker genes for the COAD dataset include CLCA1, CD8A, PRF1, GZMA, and GZMK.

IDC (Janesick et al., 2023): Invasive ductal carcinoma. This dataset includes four samples from patients with invasive ductal carcinoma. The spatial transcriptomic data are obtained using Xenium technology, comprising 44,974 transcriptomic spots and the expression levels of 541 genes. After filtering out genes with different names and blank-named genes, 500 genes are included in our study. The marker genes for the IDC dataset include MKI67, ERBB2, PRF1, ESR1, PGR, GATA3, AR, and EGFR.

PSC (Andrews et al., 2024): Primary sclerosing cholangitis. This dataset includes four samples from patients with primary sclerosing cholangitis. The spatial transcriptomic data are obtained using Visium technology, comprising 19,968 transcriptomic spots and the expression levels of 36,601 genes. Using Scanpy, the top 1000 highly variable genes are selected for each sample, and 3355 genes are included in the study. The marker genes for PSC include CYP3A4, VWF, KRT7, ACTA2, and DCN.

4.2. Evaluation metrics

We use the Pearson Correlation Coefficient (PCC) to evaluate our method, which is the most common metric for gene expression prediction (Jia et al., 2024; Pang et al., 2021; Xie et al., 2024). The PCC reflects the linear relationship between the predicted values and the true gene expression levels, with a range of $[-1, 1]$. In spatial transcriptomic gene prediction, PCC reflects the consistency between the prediction and ground truth in spatial variation patterns. The PCC is calculated as follows:

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \tag{13}$$

X and Y represent the two variables involved in the computation, i denotes the index of the data points, and n represents the total number of data points. $\overline{X}$ and $\overline{Y}$ denote the mean values of the two variables, respectively.

4.3. Implementation details

Our framework is implemented using PyTorch. Model training and evaluation are performed on a workstation equipped with an Nvidia RTX A5000 24 G GPU. All WSIs are divided into patches with a resolution of 224×224. We preprocess the gene expression data using total counts normalization and log transformation to eliminate the effects of sequencing depth. During the training phase, data augmentation techniques include random horizontal flipping, random vertical flipping, and random rotation. We use the AdamW optimizer (Loshchilov and Hutter, 2017) with a learning rate of 0.0001 to train our model for 60 epochs, with a batch size of 64.

4.4. Experimental results

We compare CMRCNet with STNet (He et al., 2020a), HisToGene (Pang et al., 2021), the latest contrastive learning-based method, BLEEP (Xie et al., 2024) and mclSTExp (Min et al., 2024). Additionally, we also compare CMRCNet with cTranPath (Wang et al., 2022a), a pioneering work of the histopathology foundation model, whose pre-training datasets share the most similar cancer groups to the datasets we used. To avoid the impact of samples with different prediction difficulties on the method's evaluation, we employ k-fold cross-validation. The number of folds is equal to the number of samples in the dataset, such that each sample participates at some point as a test sample in the evaluation of the method.

4.4.1. Comparison of highly variable genes, highly expressed genes, and marker genes

Table 1 presents the prediction results for the top 50 highly variable genes (HVGs), top 50 highly expressed genes (HEGs) among HVGs, and marker genes (MGs) used for analyzing disease progression across six datasets. The results are evaluated by PCC and Spearman's rank correlation coefficient. The contrastive learning-based methods BLEEP and CMRCNet-50 are used, with 50 similar embeddings selected. For CMRCNet-500, 500 similar embeddings are selected. On all datasets, CMRCNet demonstrates optimal or second-best results.

For the prediction of HVGs and HEGs, CMRCNet-50 shows the most significant improvement on the COAD dataset. Compared to the end-to-end prediction model STNet, CMRCNet-50 improves HEGs prediction by 25.71 % (PCC: 0.3872 vs 0.3080) and HVGs prediction by 24.12 % (PCC: 0.4611 vs 0.3715). Compared to BLEEP, CMRCNet-50 shows an improvement of 10.34 % (PCC: 0.3872 vs 0.3509) for HEGs and 10.18 % (PCC: 0.4611 vs 0.4185) for HVGs. For the SKCM dataset, CMRCNet-50 achieves PCC values of 0.3420 for HEGs and 0.1966 for HVGs, representing the current best performance.

In the prediction of marker genes, CMRCNet-50 shows a significant improvement. Compared to BLEEP, CMRCNet-50 achieves a 40.76 % improvement (PCC: 0.2704 vs 0.1921), and compared to STNet, it achieves a 330.57 % improvement (PCC: 0.2704 vs 0.0628). The results of marker genes are shown in Supplementary Tables 1–6. Fig. 3 presents box plots of the PCC scores for the top 50 HEGs and HVGs across the six datasets. The average and median PCC scores of CMRCNet-50 are significantly better than those of STNet and HisToGene, with positive PCC scores for most genes. Compared to BLEEP and mclSTExp, CMRCNet-50 achieves higher median PCC scores across most samples,

Table 1
Results of comparison experiment (top 50 genes, Multi-fold Cross-Validation). CMRCNet-50 selects 50 similar embeddings as keys for queries in the inference stage. CMRCNet-500 selects 500 similar embeddings as keys for queries in the inference stage. Best results are shown in red.

Datasets	Models	PCC			Spearman correlation		
		HEG	HVG	Marker genes	HEG	HVG	Marker genes
IDC-LymphNode	STNet	0.2067	0.1806	0.0853	0.1293	0.1464	0.0801
	HisToGene	0.2454	0.1922	0.0867	0.1623	0.1624	0.1019
	BLEEP	0.2983	0.2179	0.1076	0.1478	0.1470	0.1060
	mclSTExp	0.3038	0.2235	0.1038	0.1532	0.1521	0.1002
	CMRCNet-50	0.3156	0.2345	0.1137	0.1583	0.1599	0.1073
	CMRCNet-500	0.3370	0.2479	0.1154	0.1571	0.1535	0.1065
SKCM	STNet	0.1501	0.0825	0.0628	0.1139	0.0733	0.0770
	HisToGene	0.2359	0.1319	0.0960	0.2011	0.1096	0.1216
	BLEEP	0.3048	0.1809	0.1921	0.2664	0.1549	0.2092
	mclSTExp	0.3322	0.1754	0.1977	0.2956	0.1518	0.1948
	CMRCNet-50	0.3420	0.1966	0.2704	0.2882	0.1616	0.2611
	CMRCNet-500	0.3646	0.2268	0.2419	0.3052	0.1861	0.2354
READ	STNet	0.2257	0.2538	0.1251	0.2227	0.2403	0.0998
	HisToGene	0.1884	0.2136	0.1070	0.1798	0.1956	0.0890
	BLEEP	0.3345	0.4222	0.2541	0.3217	0.4019	0.2106
	mclSTExp	0.3372	0.4230	0.2533	0.3272	0.4028	0.2075
	CMRCNet-50	0.3327	0.4191	0.2505	0.3246	0.4014	0.2060
	CMRCNet-500	0.3361	0.4317	0.2475	0.3293	0.4158	0.2042
COAD	STNet	0.3080	0.3715	0.2221	0.3042	0.3568	0.2080
	HisToGene	0.2905	0.3438	0.1046	0.2912	0.3356	0.0911
	BLEEP	0.3509	0.4185	0.1454	0.3628	0.4288	0.1333
	mclSTExp	0.3759	0.4351	0.1419	0.3850	0.4342	0.1164
	CMRCNet-50	0.3872	0.4611	0.1776	0.3912	0.4550	0.1634
	CMRCNet-500	0.4478	0.5274	0.2136	0.4498	0.5136	0.2016
IDC	STNet	0.2520	0.2204	0.2511	0.2329	0.2225	0.2419
	HisToGene	0.2492	0.2210	0.2772	0.2336	0.2241	0.2851
	BLEEP	0.2877	0.2424	0.2869	0.2482	0.2312	0.2670
	mclSTExp	0.2732	0.2287	0.2727	0.2399	0.2198	0.2618
	CMRCNet-50	0.2993	0.2653	0.3108	0.2668	0.2576	0.3038
	CMRCNet-500	0.3174	0.2838	0.3149	0.2790	0.2774	0.3090
PSC	STNet	0.1817	0.2062	0.2382	0.1620	0.1738	0.1903
	HisToGene	0.1757	0.1967	0.1489	0.1131	0.1294	0.1506
	BLEEP	0.1837	0.2262	0.2934	0.1396	0.1647	0.2260
	mclSTExp	0.1860	0.2283	0.3051	0.1458	0.1689	0.2313
	CMRCNet-50	0.1937	0.2352	0.3002	0.1471	0.1718	0.2303
	CMRCNet-500	0.2010	0.2474	0.2990	0.1563	0.1831	0.2322

demonstrating stronger gene prediction capabilities.

In comparison with PCC, Spearman correlation is less sensitive to outliers. Across all datasets, CMRCNet-50 consistently outperforms baseline models in terms of Spearman correlation, demonstrating stronger rank-based consistency between predicted and ground-truth expression. In IDC-LymphNode and SKCM, CMRCNet-50 achieves improvements over STNet, HisToGene, and BLEEP, particularly for marker genes. For READ and COAD, CMRCNet-50 maintains competitive or superior performance across HEGs, HVGs, and marker genes, with substantial gains in COAD. In IDC and PSC, CMRCNet-50 also delivers stable improvements.

To evaluate our method's ability to predict more genes, we present the prediction results for the top 150 HEGs, HVGs, and the top 250 HEGs, HVGs in Table 2. These genes may contain those that are significant for analyzing disease progression. For the prediction of the top 150 genes, we select 250 similar embeddings for the contrastive learning-based method. In the prediction of the top 250 genes, we select 500 similar embeddings. As shown in Table 2, our method achieves the best results on most datasets. Notably, all datasets show a similar trend, where predicting more genes typically results in a lower average PCC.

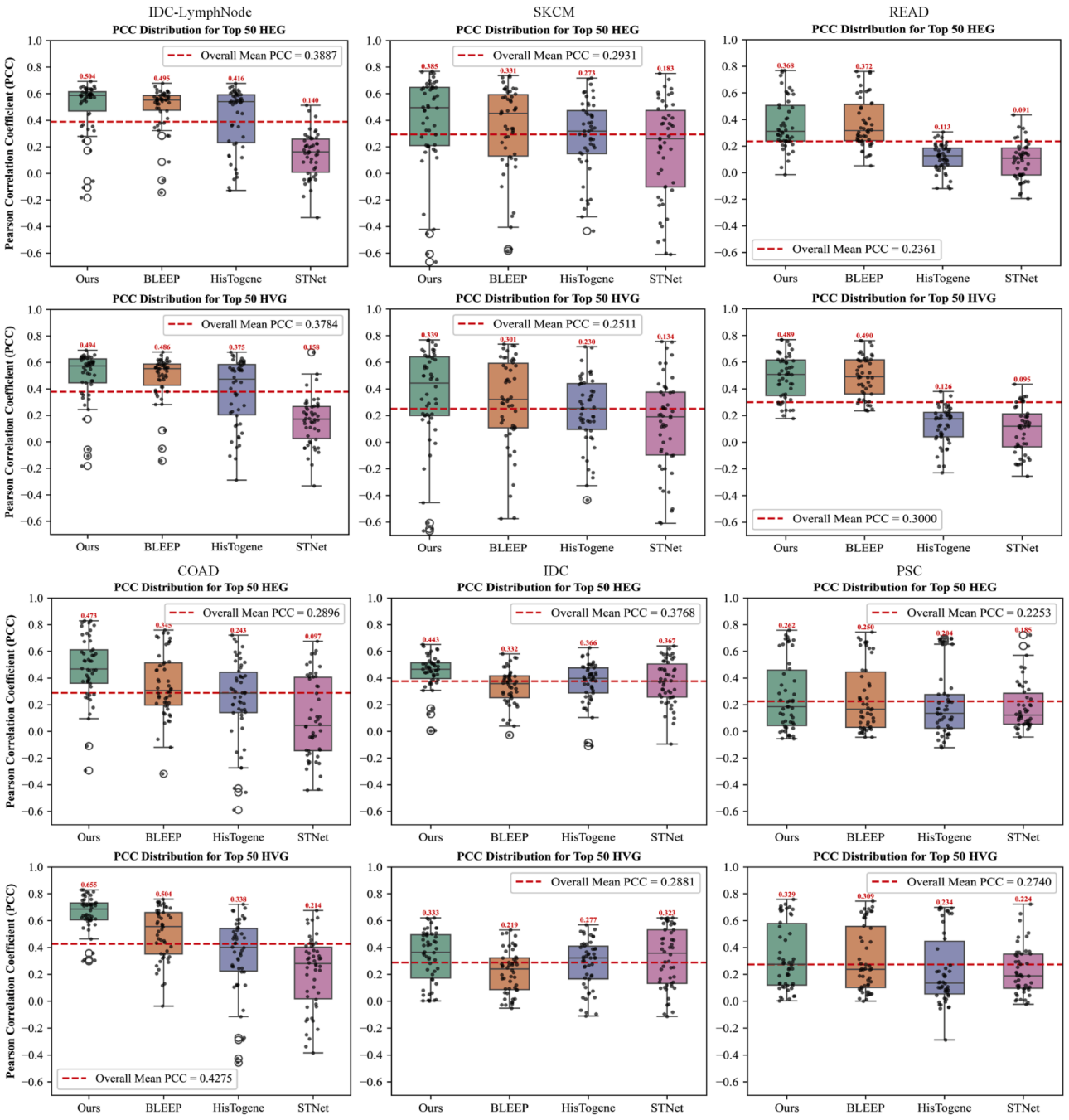

Fig. 3. Boxplot of gene PCC distribution. It reflects the predicted PCC distribution of the top 50 HEG and top 50 HVG for 6 samples from different diseases. Our method is evaluated with 50 similar embeddings. Green represents our method (CMRCNet), orange represents BLEEP, blue represents HisToGene, and pink represents STNet.

4.4.2. *Comparison of predicted and reference expression profiles*

To investigate the ability of different methods to predict gene expression levels across all genes in the samples, Fig. 4 compares the distribution of predicted expression profiles with reference expression profiles. The blue points represent the true values, while the yellow points represent the predicted values. The first row shows the average expression levels of different genes across the WSI, while the second row illustrates the variation in gene expression across different locations. In terms of predicting average expression levels, the points from STNet and HisToGene are the most scattered, indicating that it is difficult to accurately predict gene expression levels. Compared to them, BLEEP and CMRCNet-50 demonstrate excellent performance on both low-expression and high-expression genes. For genes with median expression levels, CMRCNet-50 is more robust than BLEEP. In predicting gene expression variations, CMRCNet-50 shows slight improvements over other methods but still struggles to capture variations accurately.

4.4.3. *Analysis of gene-gene correlations*

Fig. 5 shows the heatmap of gene-gene correlations, which reflects the ability of different methods to preserve biological heterogeneity. We select the top 50 HEGs and top 50 HVGs for testing. Hierarchical clustering is applied to classify and reorder these genes, with genes exhibiting similar expression patterns positioned closer together on the heatmap axes. In the heatmap of STNet's predicted results, the color blocks are scattered, indicating that the correlation between genes is not well preserved. Compared to STNet, HisToGene can recognize partial

Table 2
Results of extended genes (Multi-fold Cross-Validation). For the top 150 HEG and HVG predictions, CMRCNet selects 250 similar embeddings. For the top 250 HEG and HVG predictions, CMRCNet selects 500 similar embeddings. Best results are shown in red. All results are measured by PCC.

Datasets	Models	HEG (top 150)	HVG (top 150)	HEG (top 250)	HVG (top 250)
IDC-LymphNode	STNet	0.1891	0.1819	0.1704	0.1690
	HisToGene	0.2312	0.2131	0.2143	0.2042
	BLEEP	0.3036	0.2705	0.2892	0.2724
	mclSTExp	0.3077	0.2742	0.2936	0.2770
	CMRCNet	0.3138	0.2807	0.2983	0.2820
SKCM	STNet	0.1160	0.0971	0.0815	0.0852
	HisToGene	0.1742	0.1460	0.1365	0.1315
	BLEEP	0.2517	0.2124	0.1989	0.1961
	mclSTExp	0.2401	0.1931	0.1918	0.1899
	CMRCNet	0.2642	0.2174	0.2086	0.2054
READ	STNet	0.1631	0.1910	0.1332	0.1548
	HisToGene	0.1412	0.1637	0.1188	0.1345
	BLEEP	0.2694	0.3409	0.2326	0.2781
	mclSTExp	0.2703	0.3422	0.2330	0.2786
	CMRCNet	0.2641	0.3371	0.2268	0.2738
COAD	STNet	0.2855	0.3100	0.2610	0.2646
	HisToGene	0.2578	0.2562	0.2189	0.2173
	BLEEP	0.3510	0.3617	0.3317	0.3350
	mclSTExp	0.3465	0.3474	0.3188	0.3210
	CMRCNet	0.3709	0.3841	0.3476	0.3494
IDC	STNet	0.2275	0.2037	0.1827	0.1766
	HisToGene	0.2462	0.2064	0.1916	0.1828
	BLEEP	0.3012	0.2540	0.2526	0.2428
	mclSTExp	0.2941	0.2451	0.2461	0.2356
	CMRCNet	0.3156	0.2759	0.2678	0.2589
PSC	STNet	0.1049	0.1083	0.0754	0.0804
	HisToGene	0.0977	0.0983	0.0695	0.0726
	BLEEP	0.1235	0.1296	0.0967	0.1034
	mclSTExp	0.1240	0.1300	0.0972	0.1038
	CMRCNet	0.1288	0.1343	0.1007	0.1075

correlations between genes. Both BLEEP and our method demonstrate good performance in preserving biological heterogeneity. Regions corresponding to highly correlated gene subsets in the heatmap display similar patterns. Compared to BLEEP, our method better preserves the correlation in gene subsets that show negative correlations, represented by the blue areas in Fig. 5.

4.4.4. Analysis of highly related genes

Table 3 presents the top 5 highly correlated genes from different methods across the 6 datasets. These genes show the high PCC scores in the predicted results. In the READ (Valdeolivas et al., 2023) and PSC (Andrews et al., 2024) datasets, the most accurately predicted genes are relatively consistent across different methods. In the LYMPH_IDC (Liu et al., 2022), SKCM (Jaume et al., 2025), and COAD (Oliveira et al., 2024) datasets, some genes show consistently high correlation prediction performance. Compared to STNet (He et al., 2020a), HisToGene (Pang et al., 2021), and BLEEP (Xie et al., 2024), CMRCNet-50 shows a higher correlation with the true values for the prediction of these genes.

Figs. 6–8 show the distribution of expression levels for ITLN1, CD24, and GATA3 in colon adenocarcinoma, rectal adenocarcinoma tissue slides, and invasive ductal carcinoma tissue slides, respectively. Among them, ITLN1 (Chen et al., 2021) is an independent favorable prognostic factor in colon adenocarcinoma, with multiple functions, including a role in inhibiting tumor angiogenesis. CD24 (Huang et al., 2021) can be used to identify colorectal cancer stem cells, which are associated with tumor formation, chemotherapy resistance, and metastatic disease. GATA3 is a transcription factor that can be detected in cell nuclei of normal and neoplastic epithelial cells in the breast, and it is used as a lineage- and differentiation marker in IDC. High nuclear GATA3 expression has been shown to be associated with a favorable prognosis in certain subtypes of breast cancer. For ITLN1 prediction, BLEEP and CMRCNet-50 show higher similarity with the true values, with CMRCNet-50 achieving better PCC values. For CD24 prediction, STNet and BLEEP only show high similarity in localized regions, while HisToGene performs better in low-expression regions but struggles with predicting high-expression regions of CD24. In comparison, CMRCNet-50 better captures the spatial variation of CD24 expression levels. For the prediction of GATA3, all four methods achieve comparable results. STNet and HisToGene produce smoother expression profiles. The contrast between the prediction results of BLEEP and our

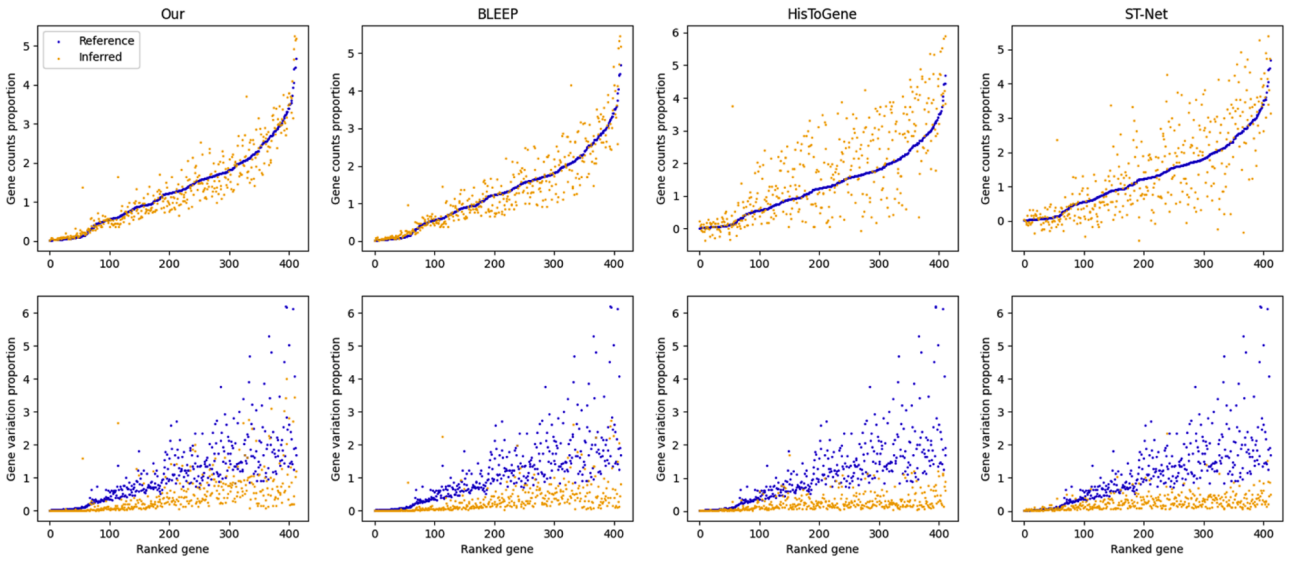

Fig. 4. Comparison of predicted and reference expression profiles. The blue points represent the distribution of ground truth and the yellow points represent the distribution of predictions. The first row represents the average expression of all genes in WSI. The second row represents the variation of all genes at different positions. Our method is evaluated with 50 similar embeddings.

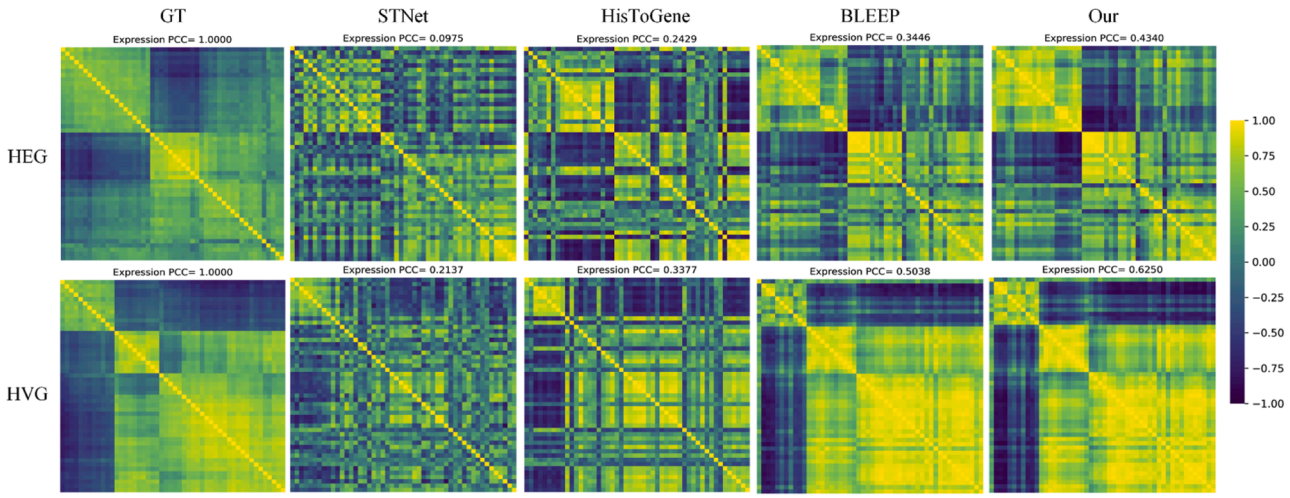

Fig. 5. Heatmap of gene-gene correlations. The first row represents the gene-gene correlations of the top 50 HEGs. The second row represents gene-gene correlations of the top 50 HVGs. Our method is evaluated with 50 similar embeddings.

method in the region of different expression levels is more obvious. The original WSIs of these samples are shown in Supplementary Figs. 1–3. Predicted PCCs of the top 50 HEGs and HVGs in these samples are shown in Supplementary Figs. 4–6.

4.4.5. Comparison with pathology foundation model

Table 4 shows the comparison between CMRCNet and cTransPath. In the LYMPH_IDC, COAD, and IDC datasets, cTransPath shows better performance, consistent with the distribution of cancer types in the cTransPath training data. For the SKCM and READ datasets, CMRCNet achieves better prediction results. Even though cTransPath's pretrained data includes relevant disease samples, the inevitable class imbalance limits cTransPath's fine-tuning ability on the relevant datasets. For the PSC dataset, cTransPath does not show the ability to predict spatial transcriptomic expression. This is because there are no cancer samples in PSC, but liver inflammation data. Even though cTransPath's pretrained data includes over 500 liver WSIs, it still struggles to bridge the gap between different disease phenotypic patterns. In contrast, CMRCNet is suitable for different disease types and achieves better performance on the PSC dataset.

4.5. Ablation experiment analysis

To demonstrate the effectiveness of our method and the impact of its components, the ablation study results are presented in Table 5. We evaluate two types of reconstruction loss functions: cosine loss and MSE loss, three different masking rates: 0.3, 0.5, 0.7, and two types of reconstructed features. The number of similarity-based embeddings used during the inference stage is set to 50. Multi-fold cross-validation is conducted, and the results are reported as the average PCC scores across six datasets. When gene expression is predicted using contrastive learning alone, the PCC scores for HEG and HVG are 0.2933 and 0.2846, respectively. When image features are reconstructed using cosine loss as the reconstruction loss, the PCC scores under different masking rates are 0.3: (HEG: 0.3090, HVG: 0.3001), 0.5: (HEG: 0.3077, HVG: 0.2968), 0.7: (HEG: 0.3063, HVG: 0.2941). When the spatial transcriptomic expression is used as the reconstructed feature, the results are HEG: 0.3069, and HVG: 0.2946. When image features are reconstructed using MSE

Table 3
Results of the top 5 correlated genes. Red: Genes showing a high correlation in all four methods. Blue: Genes showing a high correlation in three methods.

	STNet		HisToGene		BLEEP		CMRCNet	
Datasets	Gene	PCC	Gene	PCC	Gene	PCC	Gene	PCC
IDC-LymphNode	FABP4	0.4154	HSP90AA1	0.4029	UBA52	0.4851	UBA52	0.5168
	PLIN1	0.3570	FTL	0.3828	TMA7	0.4652	MYL6	0.4804
	HSP90AA1	0.3499	MYL6	0.3809	MYL6	0.4475	TMA7	0.4778
	KIAA1324	0.3480	UBA52	0.3757	HMGN2	0.4350	SRSF3	0.4491
	PLIN4	0.3449	PPIB	0.3635	SRSF3	0.4337	HMGN2	0.4454
SKCM	MLANA	0.5956	MKI67	0.5972	MKI67	0.6668	MKI67	0.6845
	S100A13	0.5880	NSG1	0.5797	SLC25A39	0.6414	S100A13	0.6446
	MKI67	0.5728	STRADB	0.5693	NSG1	0.6413	NSG1	0.6402
	MITF	0.5547	SFRP1	0.5335	TMEM150C	0.6374	SLC25A39	0.6393
	C1QA	0.5347	TMEM150C	0.5192	S100A13	0.6314	STRADB	0.6260
READ	ZG16	0.4074	ITLN1	0.3499	FCGBP	0.5361	FCGBP	0.5375
	ITLN1	0.3963	COL1A2	0.3427	MUC2	0.5105	MUC2	0.5103
	MGP	0.3900	MGP	0.3122	SPINK4	0.4976	ITLN1	0.5008
	SPINK4	0.3876	SPINK4	0.3075	ITLN1	0.4970	SPINK4	0.4914
	MALAT1	0.3753	MYL9	0.3042	ZG16	0.4913	PIGR	0.4879
COAD	DPYSL3	0.7396	CD24	0.6798	CD24	0.6488	CD24	0.6945
	MYH14	0.6580	DPYSL3	0.6243	TAGLN	0.6224	TAGLN	0.6607
	CEACAM5	0.6347	MYH14	0.6067	DPYSL3	0.6106	DPYSL3	0.6441
	IGFBP7	0.6305	SOX9	0.5958	KRTCAP3	0.5874	KRTCAP3	0.6306
	RNF43	0.6247	CEACAM5	0.5935	SCNN1A	0.5764	PPP1R1B	0.6173
IDC	MMP2	0.4582	SLC5A6	0.4647	TFAP2A	0.4799	LYPD3	0.4721
	TMEM147	0.4578	OCIAD2	0.4625	OCIAD2	0.4799	FOXA1	0.4664
	CD163	0.4328	TMEM147	0.4381	SEC24A	0.4779	TRAF4	0.4589
	LYPD3	0.4291	SRPK1	0.4291	TRAF4	0.4684	SEC11C	0.4558
	TACSTD2	0.4180	BASP1	0.4067	SLC5A6	0.4671	SEC24A	0.4533
PSC	CYP3A4	0.6351	CYP1A2	0.5351	CYP3A4	0.6540	CYP3A4	0.6551
	CYP1A2	0.5482	MT-CO3	0.5316	CYP1A2	0.5881	CYP1A2	0.5917
	ORM1	0.5075	CYP3A4	0.5225	GLUL	0.5505	GLUL	0.5470
	GLUL	0.4999	MT-CO1	0.5222	MT-CO3	0.5116	MT-CO3	0.5227
	MT-CO1	0.4642	MT-CYB	0.5199	MT-CO1	0.4937	MT-CO1	0.5105

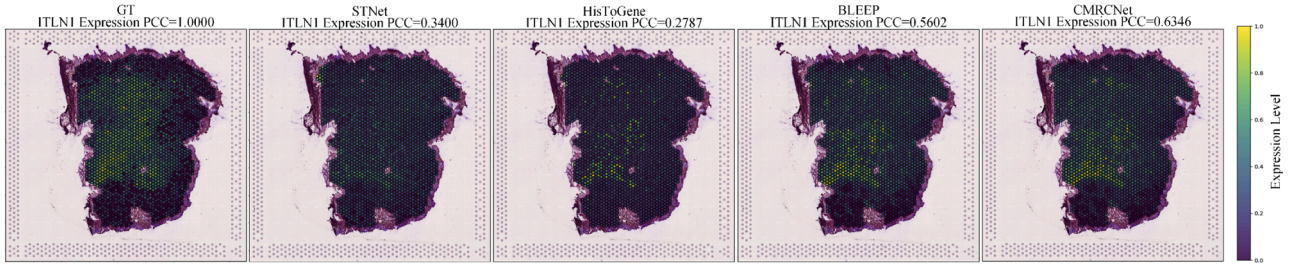

Fig. 6. Distribution of the ITLN1 in rectal cancer tissue slices.

loss as the reconstruction loss, the PCC scores under different masking rates are 0.3: (HEG: 0.3020, HVG: 0.2921), 0.5: (HEG: 0.3117, HVG: 0.3020), 0.7: (HEG: 0.3006, HVG: 0.2924). When the spatial transcriptomic expression is used as the reconstructed feature, the results are HEG: 0.2996, and HVG: 0.2937. These results consistently outperform the PCC scores obtained using contrastive learning alone.

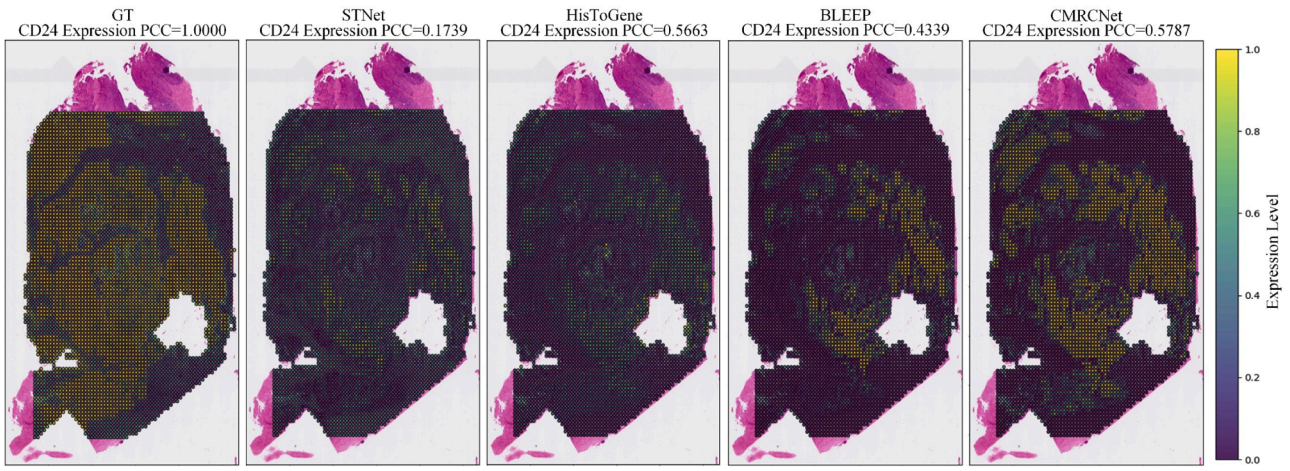

Fig. 7. Distribution of the CD24 in colon cancer tissue slices.

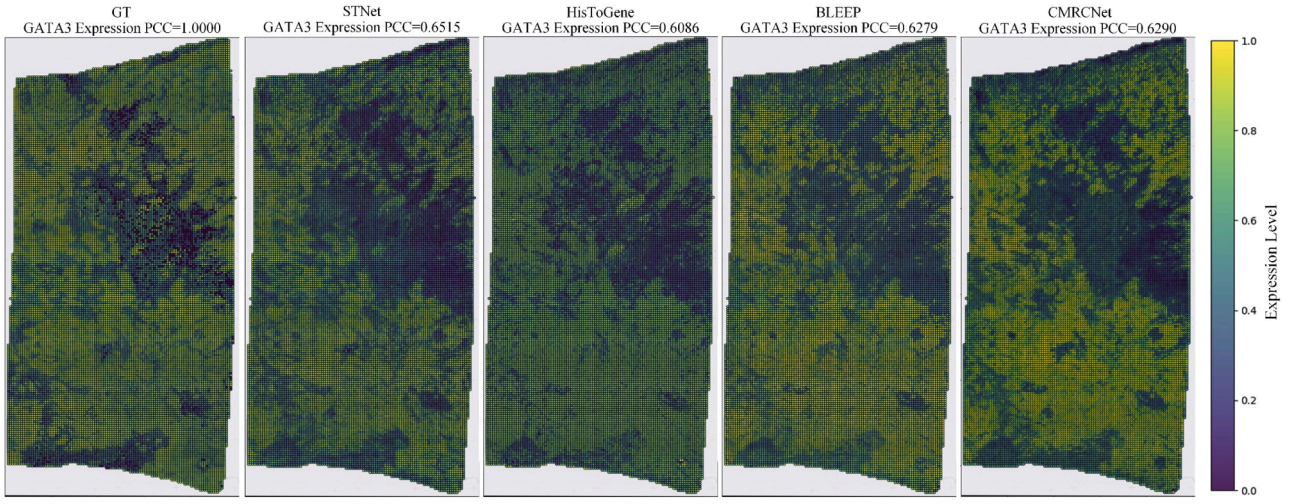

Fig. 8. Distribution of the GATA3 in breast cancer tissue slices.

Table 4
Results of comparison with pathology foundation model. Best results are shown in red.

Datasets	Models	HEG	HVG
IDC-LymphNode	cTranPath	0.4326	0.3729
	CMRCNet-500	0.3370	0.2479
SKCM	cTranPath	0.3019	0.1841
	CMRCNet-500	0.3646	0.2268
READ	cTranPath	0.1892	0.3379
	CMRCNet-500	0.3361	0.4317
COAD	cTranPath	0.4982	0.6478
	CMRCNet-500	0.4478	0.5274
IDC	cTranPath	0.4562	0.5139
	CMRCNet-500	0.3174	0.2838
PSC	cTranPath	0.0169	0.0791
	CMRCNet-500	0.2010	0.2474

Table 5
Results of ablation experiments.

Loss Function	0.3	0.5	0.7	Residual	Reconstruction	HEG	HVG	Marker Genes
						0.2933	0.2846	0.2132
cosine	√			√	Image	0.3090	0.3001	0.2263
cosine		√		√	Image	0.3077	0.2968	0.2205
cosine			√	√	Image	0.3063	0.2941	0.2282
cosine		√		√	Spatial Transcription	0.3069	0.2946	0.2116
MSE	√			√	Image	0.3020	0.2921	0.2221
MSE		√		√	Image	0.3117	0.3020	0.2372
MSE			√	√	Image	0.3006	0.2924	0.2272
MSE		√		√	Spatial Transcription	0.2996	0.2937	0.2190
MSE		√			Image	0.3042	0.2918	0.2103

Table 6
Ablation study on temperature in contrastive learning.

Temperature (τ)	HEG	HVG	Marker Genes
0.1	0.0395	0.0439	0.0194
0.5	0.2493	0.2214	0.1659
1.0	0.3117	0.3020	0.2372
2.0	0.3162	0.3031	0.2292

Table 6 reports the effect of the temperature parameter on contrastive learning performance. We observe that higher temperatures are more beneficial for morphology–spatial gene expression contrastive learning, which is consistent with previous studies (Min et al., 2024; Xie et al., 2024). At a temperature of 1.0, the PCC reaches 0.3117 for HEGs, 0.3020 for HVGs, and 0.2372 for Marker Genes. When the temperature increases to 2.0, HEG and HVG PCCs show slight improvements, whereas Marker Genes decline marginally. By contrast, lowering the temperature to 0.5 produces a marked decrease across all gene categories, and training collapses entirely at a temperature of 0.1. The role of temperature in contrastive learning is to balance the weighting of hard negatives. Smaller values enforce sharper separation between positives and their most similar negatives, while larger values encourage broader use of negative samples. In our setting, higher temperatures prove advantageous for two reasons: (1) unlike conventional image–text settings, the relationship between tissue morphology and molecular expression is inherently more complex. A lower temperature causes the model to overemphasize a small subset of hard negatives, limiting its ability to leverage diverse negative samples for modeling such complex associations. (2) Due to the limited size of our training dataset, we employ a relatively small batch size. With fewer samples per batch, overly low temperatures lead to insufficient utilization of negative samples. Conversely, large-batch training naturally provides abundant negatives, where higher temperatures encourage the model to focus on harder samples, making it more effective in data-rich scenarios.

To illustrate the role of the residual cross-attention layer, we report the impact of the residual connection on performance in the last row of

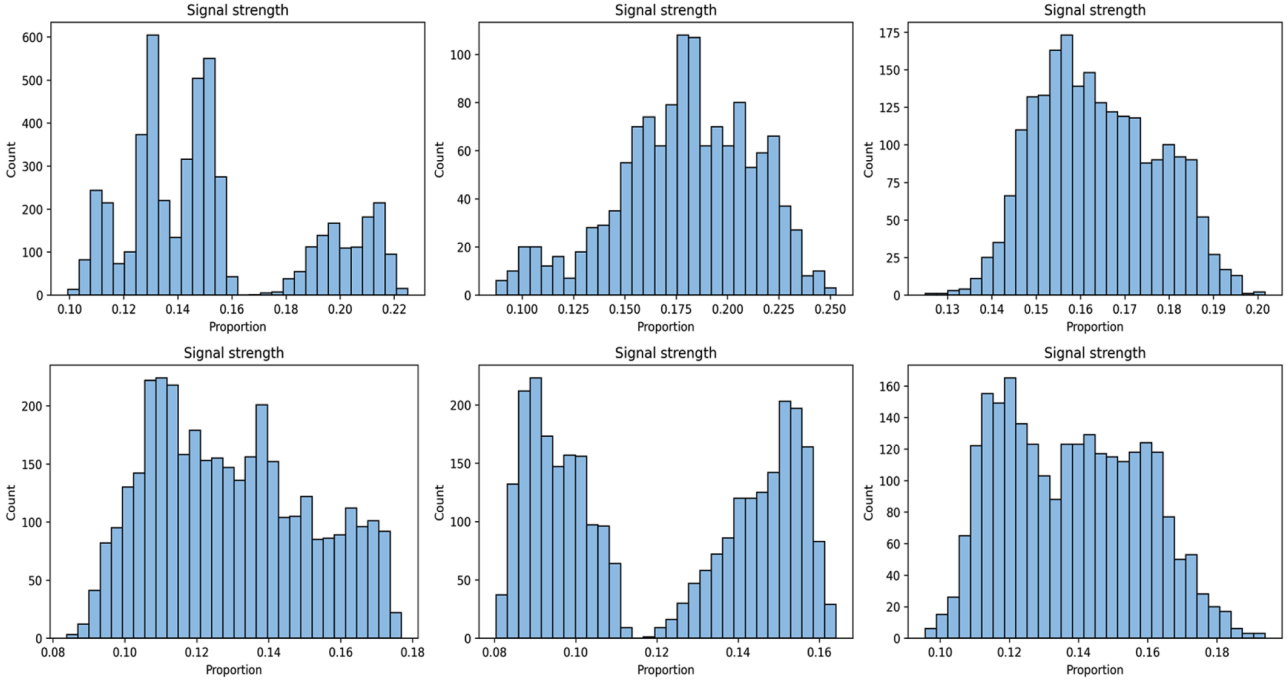

Fig. 9. Signal strength analysis of residual connections in the residual cross-attention module. The y-axis represents the number of patches, while the x-axis represents the signal strength contribution of features before the residual connection to the final output.

Table 5. When the original cross-attention mechanism is employed, the PCC values for HEGs, HVGs, and marker genes are 0.3042, 0.2918, and 0.2103, respectively. In contrast, with our residual cross-attention layer, the PCC values increase to 0.3117 for HEGs, 0.3020 for HVGs, and 0.2372 for marker genes. In addition, we analyze the signal strength before and after applying the residual connection in the second residual cross-attention module (Fig. 9). The x-axis represents the proportion of pre-residual features contributing to the final output, while the y-axis denotes the number of patches. We observe that across different datasets, most signal strengths distribute between 0.1 and 0.2, which indicates that the residual connection contributes the majority of the signal strength to the final output. In our design, signals from histopathology images contribute to cross-modal reconstruction through the residual connection (as illustrated in Fig. 2). On the one hand, this design allows the reconstruction process to benefit not only from gene features but also from the continuity of image-derived features, which facilitates the self-reconstruction. On the other hand, the residual connection facilitates training and alleviates the vanishing gradient problem.

5. Discussion

In this work, we propose a novel method (CMRCNet), which is based on contrastive learning, for predicting spatially resolved gene expression detected by spatial transcriptomics from routinely stained pathology images. During the training phase, contrastive learning is employed to minimize the distance between patches and their paired spatial transcriptomic expression in feature space. In the inference phase, CMRCNet selects gene expression embeddings that are similar in feature space as keys for each patch, and the mean of the original expression of these keys is computed as the prediction. Experiments on datasets from six different types of diseases demonstrate that CMRCNet outperforms others in spatial transcriptomic expression prediction. It not only predicts gene spatial expression trends more accurately but also shows advantages in preserving biological heterogeneity.

CLIP (Radford et al., 2021) demonstrates remarkable potential in the image-text tasks, providing a paradigm for cross-modal data alignment. BLEEP is the first to adapt CLIP for spatial transcriptomic expression prediction by aligning histological images with spatial transcriptomics data through contrastive learning. However, previous studies show that contrastive learning, which relies solely on similarity-based objectives, struggles to bridge the gap between modalities (Li et al., 2021; Zhang et al., 2024). Numerous works propose various auxiliary tasks to further narrow the distance between images and text in semantic space. For instance, ALBEP (Li et al., 2021) selects hard samples from a mini-batch and designs a classification task to optimize semantic matching. CE-CLIP (Zhang et al., 2024) enhances CLIP's inference in visual-text combinations by directly generating hard samples. However, these methods are difficult to transfer to histopathology images and spatial transcriptomics data for two main reasons. First, text data contains rich and well-defined semantic information. Nouns in text represent the subjects in images, and adjectives reveal the characteristics of these subjects. For this reason, these methods (Li et al., 2021; Zhang et al., 2024) optimize the semantic matching process of image-text pairs. In contrast, spatial transcriptomic data lack explicit semantic information, and the associations between different genes are complex. Moreover, noise is prevalent in the data, which affects the modeling of image-gene pairs. Second, a large number of image-text pairs are readily available for model training. CLIP is trained on 400 million image-text pairs. Thanks to the rich semantic information and vast training data, CLIP achieves success in the image and text domains. However, this success is difficult to replicate in the context of pathology patches and spatial transcriptomics. Collecting a large amount of spatial transcriptomic data for training is challenging, and further limitations arise from the different disease representation patterns and gene lengths.

We propose CMRCNet, an improved version of CLIP tailored for histopathology images and spatial transcriptomics, designed to predict gene expression from images. Unlike semantic-driven methods, our method operates directly at the feature-pixel level. It incorporates two loss functions: a contrastive learning loss and a reconstruction loss. The contrastive learning loss establishes associations between paired cross-modal data in the feature space through similarity computation, while the reconstruction loss directly integrates multimodal data at the feature level, thereby facilitating cross-modal interaction learning. Furthermore, our approach enables training and inference on small-scale datasets. One of the main reasons is that it provides denser and more fixed supervision signals. Specifically, a single sample can generate multiple local prediction tasks by reconstructing the masked features, which substantially increases the amount of effective supervision. In addition, cross-modal reconstruction can be regarded as a supervised task with a fixed objective. Such a fixed target avoids the noise and instability observed in other approaches—such as semantic matching or hard negative mining—where the effectiveness heavily depends on sufficient samples and the appropriate selection of negatives. In other words, this stable supervision signal serves as an implicit regularization mechanism, enhancing contrastive learning stability on small-scale datasets.

Compared to end-to-end prediction models, it offers greater flexibility during inference, as it does not require fixed labels. Most end-to-end methods focus on predicting a fixed number of highly variable genes within a dataset (He et al., 2020a; Jia et al., 2024; Pang et al., 2021). However, due to expression heterogeneity across samples and the fact that certain marker genes do not exhibit strong variability, these important genes are often overlooked. Our method accounts for sample-specific differences during inference, resulting in greater generalizability. BLEEP is the method most closely related to ours, as it is also based on the contrastive learning paradigm. However, BLEEP does not provide further insights into how to enhance the performance of histology–molecular expression contrastive learning. In comparison, our method leverages cross-modal reconstruction to further reduce the distance between images and molecular expressions in the feature space. To further analyze the predictive capability of CMRCNet, we compare CMRCNet with the pathology foundation model cTransPath. cTransPath is trained on over 35,000 WSIs, totaling approximately 15 million patches, and includes multiple cancer types. The most common tissue types in these datasets are breast tissue, colon tissue, kidney tissue, and lung tissue, as common cancers often occur in these tissues. Compared to cTransPath, our image encoder uses the VIT-B version, which is trained only on the datasets we used, without incorporating any pretrained knowledge. In addition, we conduct more comprehensive and reliable validation. By evaluating six different disease datasets and employing multi-fold cross-validation, we mitigate the randomness introduced by varying levels of prediction difficulty across samples within the datasets.

In the ablation study, the impact of the cross-modal reconstruction loss is highly significant. Overall, across different experimental settings, the results of CMRCNet consistently outperform models that rely solely on contrastive learning loss. Specifically, when using the cosine similarity loss for model optimization, we recommend selecting a lower mask rate. Cosine similarity loss measures the similarity between feature vectors in the embedding space, and preserving more original information is beneficial for model learning. When optimizing the model with MSE loss, an appropriate mask rate is more critical, which in our task is 0.5. MSE focuses more on feature pixels, where a smaller masking ratio may fail to fully activate the model's potential, while a higher ratio may introduce learning difficulties. Additionally, under the same loss function and mask rate, reconstructing image features proves to be the better choice. The superior performance of reconstructing image features arises from the fact that gene expression exerts a more direct influence on morphology. From a biological perspective, genes interact with each other and regulate cellular behavior through biological pathways, which in turn lead to morphological changes. Therefore, gene expression can be more easily mapped to morphology. In contrast,

image features are more complex, and mapping them back to gene expression is inherently more challenging. In our setup, image features are chosen as the reconstruction target rather than the original histological images, which differs from the MAE algorithm that inspired our approach. The rationale behind this design is twofold. First, our goal is to minimize the distance between image and gene representations in the feature space, rather than to reconstruct or generate the original images. Using image features as the reconstruction target offers the advantage that, during the cross-modal reconstruction stage, both morphological and gene features are explicitly enforced to be encoded within the same latent space, which aligns with our objective. In contrast, if the original images were used as the reconstruction target, the image encoder would tend to retain low-level information beneficial for image reconstruction, rather than learning a shared multimodal representation space. Second, histopathology images are highly complex, with each patch containing rich structural details, primarily multiple cells and their morphologies. Reconstructing such details requires a powerful decoder, the optimization of which is not easy, particularly when sufficient training data are unavailable. This makes raw image reconstruction extremely challenging.

A major limitation of our method lies in inference speed, which stems from its unique design paradigm. End-to-end models directly infer gene expression levels from histopathology images. In contrast, our method requires selecting similar transcriptomic embeddings for each patch in the test cohort. Consequently, the computation of the similarity matrix involves all samples in the dataset. As a result, in addition to being influenced by the complexity of the encoder, the inference speed is also dependent on the number of samples in the dataset. Another major limitation is the potential for high variance, which arises from two main factors. (1) The inclusion of larger-scale datasets is challenging, as spatial transcriptomics data are scarce. Limited sample sizes often fail to capture the diversity of the true distribution, leading to models trained under such conditions being more uncertain. (2) The inherent biological heterogeneity across individuals. From the biological perspective, even patients with the same disease may display substantial differences in tissue morphology and gene expression profiles. At the experimental level, model performance may vary when different samples are used for training and testing. Such heterogeneity can inevitably contribute to high variance. For the first limitation, optimizing the inference process in real clinical practice is of critical importance, such as employing mixed-precision inference and optimizing memory copy operations. For the second limitation, strengthening collaboration with the clinical community to collect large-scale spatial transcriptomics datasets represents the most direct solution. In addition, the heterogeneity across samples should be further evaluated by biological experts.

Future work can be explored in two directions. (1) Our method has the potential to be transferred to other tasks. In this study, CMRCNet demonstrates promising performance in predicting marker genes, as shown in Figs. 6–8. The expression of these genes can be used to distinguish relevant parts of cancerous tissues from normal tissues (e.g., presence of CD27 or absence of ITLN1), or to identify cells with a particular type of differentiation or histogenetic “lineage” (e.g., GATA3). Furthermore, we establish correlations between histopathology images and transcriptomic data through contrastive learning. A multimodal knowledge-embedded encoder can leverage knowledge distillation techniques to transfer these correlations to tasks such as survival prediction and cancer subtype classification, thereby enhancing performance under a single modality. (2) Our method can be extended to a wide range of different markers in other data sets and may also be used in other types of spatially resolved image data. Future applications of the method in real-world clinical settings may include computer-based prediction of biomarker expression on routinely stained tissues to guide subsequent biomarker analysis, for example, to select appropriate marker panels, or to save time and accelerate case management (e.g., initiate clinical staging) even before the more time-consuming standard biomarker analysis is completed. AI–based diagnosis has already been validated across multiple data modalities, such as fundus images, CT, and MRI (Shin, 2024). However, its evaluation on spatial transcriptomics data remains insufficient. In addition, since the introduction of CLIP, numerous CLIP-based methods have been developed to model the correlations between medical images and clinical text reports. However, very few universal methods have been proposed for structured clinical data, such as spatial transcriptomic expression, radiomic features, and serum antigen expression levels. In the future, we will validate our model on further types of medical data.

6. Conclusion

In this study, we propose CMRCNet, a novel method for inferring spatial gene expression levels from histopathological morphology. CMRCNet leverages contrastive learning to align morphological features with spatial transcriptomics in the latent space and generates spatial gene expression based on cross-modal feature similarity. To enhance the standard contrastive learning framework, we introduce a cross-modal masked reconstruction as a pretext task. Unlike common contrastive learning approaches that rely heavily on large-scale datasets and semantically rich data, CMRCNet is well-suited for the structured data commonly encountered in medical applications. Comprehensive quantitative and qualitative analyses across six datasets from different organs demonstrate that CMRCNet not only achieves superior performance in predicting highly expressed and highly variable genes, but also effectively preserves gene–gene correlations. Furthermore, it shows strong potential in supporting clinical WSI analysis by accurately localizing tumor regions or identifying cells with a particular type of differentiation based on biomarker gene expression.

CRediT authorship contribution statement

Junzhuo Liu: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Formal analysis, Data curation, Conceptualization. **Markus Eckstein:** Writing – review & editing. **Zhixiang Wang:** Writing – review & editing. **Friedrich Feuerhake:** Writing – review & editing, Supervision, Project administration, Conceptualization. **Dorit Merhof:** Writing – review & editing, Supervision, Project administration, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research is partially funded by the China Scholarship Council (CSC).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.media.2025.103889.

References

Al-Kateb, G., Cengiz, E., Gök, M., 2025. BioGPT: a Generative Transformer-Based Framework for Personalized Genomic Medicine and Rare Disease Diagnosis. Mesop. J. Artif. Intell. Healthc. 2025, 154–164.

Andrews, T.S., Nakib, D., Perciani, C.T., Ma, X.Z., Liu, L., Winter, E., Camat, D., Chung, S. W., Lumanto, P., Manuel, J., 2024. Single-cell, single-nucleus, and spatial transcriptomics characterization of the immunological landscape in the healthy and PSC human liver. J. Hepatol 80, 730–743.

Baldrati, A., Bertini, M., Uricchio, T., Del Bimbo, A., 2022. Effective conditioned and composed image retrieval combining clip-based features, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 21466–21474.

Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R.L., Soerjomataram, I., Jemal, A., 2024. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin 74, 229–263.
Brierley, J.D., Gospodarowicz, M.K., Wittekind, C., 2017. TNM Classification of Malignant Tumours. John Wiley & Sons.
Cannella, R., Matteini, F., Dioguardi Burgio, M., Sartoris, R., Beaufrère, A., Calderaro, J., Mulé, S., Reizine, E., Luciani, A., Laurent, A., 2024. Association of LI-RADS and histopathologic features with survival in patients with solitary resected hepatocellular carcinoma. Radiology. 310, e231160.
Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A., 2020. Unsupervised learning of visual features by contrasting cluster assignments. Adv. Neural Inf. Process. Syst. 33, 9912–9924.
Chen, L., Jin, X.-H., Luo, J., Duan, J.-L., Cai, M.-Y., Chen, J.-W., Feng, Z.-H., Guo, A.M., Wang, F.-W., Xie, D., 2021. ITLN1 inhibits tumor neovascularization and myeloid derived suppressor cells accumulation in colorectal carcinoma. Oncogene 40, 5925–5937.
Chen, R.J., Lu, M.Y., Wang, J., Williamson, D.F., Rodig, S.J., Lindeman, N.I., Mahmood, F., 2020a. Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. IEEe Trans. Med. ImAging 41, 757–770.
Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020b. A simple framework for contrastive learning of visual representations. In: International conference on machine learning. PmLR, pp. 1597–1607.
Chen, W.-T., Lu, A., Craessaerts, K., Pavie, B., Frigerio, C.S., Corthout, N., Qian, X., Laláková, J., Kühnemund, M., Voytyuk, I., 2020c. Spatial transcriptomics and in situ sequencing to study Alzheimer's disease. Cell 182, 976–991 e919.
Ding, K., Zhou, M., Metaxas, D.N., Zhang, S., 2023. Pathology-and-genomics multimodal transformer for survival outcome prediction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 622–631.
Dosovitskiy, A., 2020. An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.
Ektefaie, Y., Yuan, W., Dillon, D.A., Lin, N.U., Golden, J.A., Kohane, I.S., Yu, K.-H., 2021. Integrative multiomics-histopathology analysis for breast cancer classification. NPJ. Breast. Cancer 7, 147.
Fu, Y., Jung, A.W., Torne, R.V., Gonzalez, S., Vöhringer, H., Shmatko, A., Yates, L.R., Jimenez-Linan, M., Moore, L., Gerstung, M., 2020. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. Nat. Cancer 1, 800–810.
Garcia-Alonso, L., Handfield, L.-F., Roberts, K., Nikolakopoulou, K., Fernando, R.C., Gardner, L., Woodhams, B., Arutyunyan, A., Polanski, K., Hoo, R., 2021. Mapping the temporal and spatial dynamics of the human endometrium in vivo and in vitro. Nat. Genet 53, 1698–1711.
Hamilton, W., Ying, Z., Leskovec, J., 2017. Inductive representation learning on large graphs. Adv. Neural inf. Process. Syst. 30.
Handley, K.F., Sims, T.T., Bateman, N.W., Glassman, D., Foster, K.I., Lee, S., Yao, J., Yao, H., Fellman, B.M., Liu, J., 2022. Classification of high-grade serous ovarian cancer using tumor morphologic characteristics. JAMA netw. open 5, e2236626. -e2236626.
He, B., Bergenstråhle, L., Stenbeck, L., Abid, A., Andersson, A., Borg, Å., Maaskola, J., Lundeberg, J., Zou, J., 2020a. Integrating spatial gene expression and breast tumour morphology via deep learning. Nat. Biomed. Eng. 4, 827–834.
He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R., 2022. Masked autoencoders are scalable vision learners, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 16000–16009.
He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020b Momentum contrast for unsupervised visual representation learning, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 9729–9738.
Ho, K.K., Fleseriu, M., Wass, J., Katznelson, L., Raverot, G., Little, A.S., Castaño, J.P., Reincke, M., Lopes, M.B., Kaiser, U.B., 2024. A proposed clinical classification for pituitary neoplasms to guide therapy and prognosis. Lancet Diabetes Endocrinol. 12, 209–214.
Hölscher, D.L., Bouteldja, N., Joodaki, M., Russo, M.L., Lan, Y.-C., Sadr, A.V., Cheng, M., Tesar, V., Stillfried, S.V., Klinkhammer, B.M., 2023. Next-Generation Morphometry for pathomics-data mining in histopathology. Nat. Commun. 14, 470.
Hu, J., Li, X., Coleman, K., Schroeder, A., Ma, N., Irwin, D.J., Lee, E.B., Shinohara, R.T., Li, M., 2021. SpaGCN: integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. Nat. Methods 18, 1342–1351.
Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700–4708.
Huang, J.L., Oshi, M., Endo, I., Takabe, K., 2021. Clinical relevance of stem cell surface markers CD133, CD24, and CD44 in colorectal cancer. Am. J. Cancer Res. 11, 5141.
Huang, W., Xu, M., Hu, X., Abousamra, S., Ganguly, A., Kapse, S., Yurovsky, A., Prasanna, P., Kurc, T., Saltz, J., 2024. RankByGene: gene-Guided Histopathology Representation Learning Through Cross-Modal Ranking Consistency. arXiv preprint arXiv:2411.15076.
Hunter, M.V., Moncada, R., Weiss, J.M., Yanai, I., White, R.M., 2021. Spatially resolved transcriptomics reveals the architecture of the tumor-microenvironment interface. Nat. Commun. 12, 6278.
Janesick, A., Shelansky, R., Gottscho, A.D., Wagner, F., Williams, S.R., Rouault, M., Beliakoff, G., Morrison, C.A., Oliveira, M.F., Sicherman, J.T., 2023. High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. Nat. Commun. 14, 8353.
Jaume, G., Doucet, P., Song, A., Lu, M.Y., Almagro Pérez, C., Wagner, S., Vaidya, A., Chen, R., Williamson, D., Kim, A., 2025. Hest-1k: a dataset for spatial transcriptomics and histology image analysis. Adv. Neural Inf. Process. Syst. 37, 53798–53833.
Jia, Y., Liu, J., Chen, L., Zhao, T., Wang, Y., 2024. THItoGene: a deep learning method for predicting spatial transcriptomics from histological images. Br. Bioinform. 25, bbad464.
Jin, Y., Zuo, Y., Li, G., Liu, W., Pan, Y., Fan, T., Fu, X., Yao, X., Peng, Y., 2024. Advances in spatial transcriptomics and its applications in cancer research. Mol. Cancer 23, 129.
Khaliq, A.M., Rajamohan, M., Saeed, O., Mansouri, K., Adil, A., Zhang, C., Turk, A., Carstens, J.L., House, M., Hayat, S., 2024. Spatial transcriptomic analysis of primary and metastatic pancreatic cancers highlights tumor microenvironmental heterogeneity. Nat. Genet 56, 2455–2465.
Kilim, O., Olar, A., Biricz, A., Madaras, L., Pollner, P., Szállási, Z., Sztupinszki, Z., Csabai, I., 2025. Histopathology and proteomics are synergistic for high-grade serous ovarian cancer platinum response prediction. npj Precis. Oncol. 9, 27.
Klambauer, G., Unterthiner, T., Mayr, A., Hochreiter, S., 2017. Self-normalizing neural networks. Adv. Neural Inf. Process. Syst. 30.
Kleihues, P., Louis, D.N., Scheithauer, B.W., Rorke, L.B., Reifenberger, G., Burger, P.C., Cavenee, W.K., 2002. The WHO classification of tumors of the nervous system. J. Neuropathol. Exp. Neurol. 61, 215–225.
Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H., 2021. Align before fuse: vision and language representation learning with momentum distillation. Adv. Neural Inf. Process. Syst. 34, 9694–9705.
Liu, T., Liu, C., Yan, M., Zhang, L., Zhang, J., Xiao, M., Li, Z., Wei, X., Zhang, H., 2022. Single cell profiling of primary and paired metastatic lymph node tumors in breast cancer patients. Nat. Commun. 13, 6823.
Loshchilov, I., Hutter, F., 2017. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.
Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., Li, T., 2022. Clip4clip: an empirical study of clip for end to end video clip retrieval and captioning. Neurocomputing. 508, 293–304.
Ma, Y., Xu, G., Sun, X., Yan, M., Zhang, J., Ji, R., 2022. X-clip: end-to-end multi-grained contrastive learning for video-text retrieval, Proceedings of the 30th ACM international conference on multimedia, pp. 638–647.
Mejia, G., Cárdenas, P., Ruiz, D., Castillo, A., Arbeláez, P., 2023. SEPAL: spatial gene expression prediction from local graphs, Proceedings of the IEEE/CVF International Conference on computer vision, pp. 2294–2303.
Min, W., Shi, Z., Zhang, J., Wan, J., Wang, C., 2024. Multimodal contrastive learning for spatial gene expression prediction using histology images. Br. Bioinform 25, bbae551.
Moncada, R., Barkley, D., Wagner, F., Chiodin, M., Devlin, J.C., Baron, M., Hajdu, C.H., Simeone, D.M., Yanai, I., 2020. Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. Nat. Biotechnol 38, 333–342.
Oliveira, M.F., Romero, J.P., Chung, M., Williams, S., Gottscho, A.D., Gupta, A., Pilipauskas, S.E., Mohabbat, S., Raman, N., Sukovich, D., 2024. Characterization of immune cell populations in the tumor microenvironment of colorectal cancer using high definition spatial profiling. bioRxiv., 597233, 2024.2006. 2004.
Palacio, S., Engler, P., Hees, J., Dengel, A., 2021. Contextual classification using self-supervised auxiliary models for deep neural networks. In: 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, pp. 8937–8944.
Pang, M., Su, K., Li, M., 2021. Leveraging information in spatial transcriptomics to predict super-resolution gene expression from histology images in tumors. bioRxiv., 470212, 2021.2011. 2028.
Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., 2021. Learning transferable visual models from natural language supervision. In: International conference on machine learning. PMLR, pp. 8748–8763.
Rahaman, M.M., Millar, E.K., Meijering, E., 2023. Breast cancer histopathology image-based gene expression prediction using spatial transcriptomics data and deep learning. Sci. Rep. 13, 13604.
Rakha, E., Toss, M., Quinn, C., 2022. Specific cell differentiation in breast cancer: a basis for histological classification. J. Clin. Pathol 75, 76–84.
Shin, J., 2024. Revolutionizing medical imaging with artificial intelligence real-time segmentation for enhanced diagnostics. EDRAAK 2024, 18–25.
Sparks, R., Madabhushi, A., 2013. Explicit shape descriptors: novel morphologic features for histopathology classification. Med. Image Anal. 17, 997–1009.
Ståhl, P.L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J.F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J.O., Huss, M., 2016. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science 353, 78–82.
Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F., 2021. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin 71, 209–249.
Swanton, C., Bernard, E., Abbosh, C., André, F., Auwerx, J., Balmain, A., Bar-Sagi, D., Bernards, R., Bullman, S., DeGregori, J., 2024. Embracing cancer complexity: hallmarks of systemic disease. Cell 187, 1589–1616.
Takahara, T., Murase, Y., Tsuzuki, T., 2021. Urothelial carcinoma: variant histology, molecular subtyping, and immunophenotyping significant for treatment outcomes. Pathology. 53, 56–66.
Tao, M., Bao, B.-K., Tang, H., Xu, C., 2023. Galip: generative adversarial clips for text-to-image synthesis, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 14214–14223.

Tian, Y., Krishnan, D., Isola, P., 2020. Contrastive multiview coding, Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16. Springer, pp. 776–794.
Tsuneki, M., Abe, M., Ichihara, S., Kanavati, F., 2023. Inference of core needle biopsy whole slide images requiring definitive therapy for prostate cancer. BMC. Cancer 23, 11.
Valdeolivas, A., Amberg, B., Giroud, N., Richardson, M., Gálvez, E.J., Badillo, S., Julien-Laferrière, A., Turos, D., von Voithenberg, L.V., Wells, I., 2023. Charting the heterogeneity of colorectal cancer consensus molecular subtypes using spatial transcriptomics. bioRxiv., 525135, 2023.2001. 2023.
Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., Han, X., 2022a. Transformer-based unsupervised contrastive learning for histopathological image classification. Med. Image Anal. 81, 102559.
Wang, Y., Liu, B., Min, Q., Yang, X., Yan, S., Ma, Y., Li, S., Fan, J., Wang, Y., Dong, B., 2023. Spatial transcriptomics delineates molecular features and cellular plasticity in lung adenocarcinoma progression. Cell Discov. 9, 96.
Wang, Z., Liu, W., He, Q., Wu, X., Yi, Z., 2022b Clip-gen: language-free training of a text-to-image generator with clip. arXiv preprint arXiv:2203.00386.
Williams, C.G., Lee, H.J., Asatsuma, T., Vento-Tormo, R., Haque, A., 2022. An introduction to spatial transcriptomics for biomedical research. Genome Med. 14, 68.
Wolf, F.A., Angerer, P., Theis, F.J., 2018. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 19, 1–5.
Wu, Z., Xiong, Y., Yu, S.X., Lin, D., 2018. Unsupervised feature learning via non-parametric instance discrimination, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3733–3742.
Xie, R., Pang, K., Chung, S., Perciani, C., MacParland, S., Wang, B., Bader, G., 2024. Spatially Resolved Gene Expression Prediction from Histology Images via Bi-modal Contrastive Learning. Adv. Neural Inf. Process. Syst. 36.
Yang, Y., Hossain, M.Z., Stone, E., Rahman, S., 2024. Spatial transcriptomics analysis of gene expression prediction using exemplar guided graph neural network. Pattern. Recognit. 145, 109966.
Yang, Y., Hossain, M.Z., Stone, E.A., Rahman, S., 2023. Exemplar guided deep neural network for spatial transcriptomics analysis of gene expression prediction, Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 5039–5048.
Yu, Q., Jiang, M., Wu, L., 2022. Spatial transcriptomics technology in cancer research. Front. Oncol 12, 1019111.
Zhang, L., Awal, R., Agrawal, A., 2024. Contrasting intra-modal and ranking cross-modal hard negatives to enhance visio-linguistic compositional understanding, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13774–13784.