ISG: I can See Your Gene Expression

Yan Yang[1]
u6169130@anu.edu.au
LiYuan Pan[2]
liyuan.pan@bit.edu.cn
Liu Liu[3]
nwpuliuliu@gmail.com
Eric A Stone[1†]
eric.stone@anu.edu.au

[1] Biological Data Science Institute,
Research School of Biology,
The Australian National University,
Australia
[2] BITSZ & School of CSAT, BIT, China
[3] Cyberverse Lab, China

Abstract

This paper aims to predict gene expression from a histology slide image precisely. Such a slide image has a large resolution and sparsely distributed textures. These obstruct extracting and interpreting discriminative features from the slide image for diverse gene types prediction. Existing gene expression methods mainly use general components to filter textureless regions, extract features, and aggregate features uniformly across regions. However, they ignore gaps and interactions between different image regions and are therefore inferior in the gene expression task (Sec. 1). Instead, we present **ISG** framework that harnesses interactions among discriminative features from texture-abundant regions by three new modules: 1) a *Shannon Selection* module (Sec. 3.1), based on the Shannon information content and Solomonoff's theory, to filter out textureless image regions; 2) a *Feature Extraction* network (Sec. 3.2) to extract expressive low-dimensional feature representations for efficient region interactions among a high-resolution image; 3) a *Dual Attention* network (Sec. 3.3) attends to regions with desired gene expression features and aggregates them for the prediction task. Extensive experiments on standard benchmark datasets show that the proposed **ISG** framework outperforms state-of-the-art methods significantly.

1 Introduction

Gene expression prediction from a histology slide image is an indispensable component for efficiently understanding clinic treatment developments [4, 6, 29]. A histology slide image has two characteristics: i) it has a large resolution amounting to $10^5 \times 10^5$ [29]. The large resolution prevents an end-to-end solution, i.e., directly using traditional deep learning approaches (e.g., convolution [7] and transformer [5, 33] networks), for high computational cost; ii) it has sparse and non-uniformly distributed textures (See Fig. 4) which hinder model inference [39]. To predict gene expression precisely, feature extractions and interactions of regions with different texture levels among the histology slide image need to be explored.

To date, this gene prediction problem remains under-explored. The pioneer HE2RNA [29] provides a three-stage solution. First, it tiles a histology slide image into patches and

[†] Corresponding author.

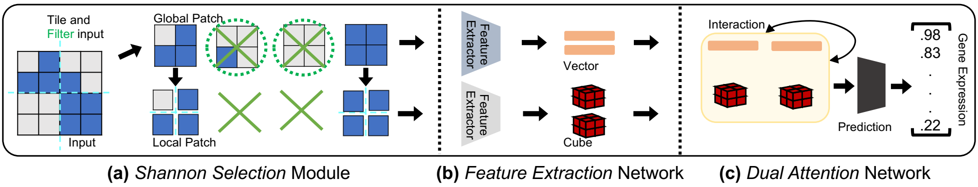


Figure 1: ***ISG** framework. **(a)** Blue and gray squares separately denote texture-abundant and textureless patches. Given an input image, it is tiled using two (coarse and fine) resolutions, resulting in global and local patches. Textureless/featureless patches are filtered out with a Shannon Selection module. **(b)** Taking global and local patches as input, two separate Feature Extraction networks are used to extract low-dimension feature representations, resulting in global and local features. Each global feature corresponds with multiple local features, or equivalently, a local feature cube. **(c)** Taking global and local features as input, our Dual Attention Network brings interactions to these two types of features before predictions.*

filters out patches with high background noise via the Otsu algorithm [23]. Second, patch features are extracted from a pretrained ResNet [7] and clustered with a K-mean algorithm by patch locations. Third, the aggregated cluster-wise features are independently processed by a multi-layer perceptron (MLP) [36] and an average pooling layer, for gene expression prediction. However, HE2RNA has three limitations: i) Textureless patches, e.g., patches with a solid colour or scattered noisy chunks (see Fig. 2), are not filtered; ii) The ResNet pre-trained on ImageNet fails to identify histology-related features due to dataset gaps (see Sec. 4.2); and iii) feature interactions among patches are not considered, which neglects long-range dependency between patches for gene expression predictions.

To address the above limitations, by analysing the characteristics of the histology slide image, we propose an **ISG** framework (see Fig. 1) that builds feature interactions between patches with abundant textures and injects global contextual information into features to make better predictions. Our **ISG** has three new modules connected in a sequence: 1) a theoretical *Shannon Selection* module (Sec. 3.1). It quantifies the patch texture abundance levels. Given an input histology slide image, it is first segmented into patches at two resolutions - coarse and fine. We separately name the ‘coarse-resolution’ and ‘fine-resolution’ patches to ‘global’ and ‘local’ patches. The *Shannon Selection* module selects patches with a large length of minimal description by incorporating Solomonoff’s universal prior and Shannon information content; 2) a *Feature Extraction* network (Sec. 3.2). Given ‘global’ and ‘local’ patches, two separate *Feature Extraction* networks are used to extract discriminative patch representations. Both fine-grained local features and coarse-level global features are obtained. This module follows an unsupervised manner, as there is a relatively large amount of images compared to the available label annotations (e.g., each gene expression label pairs with an image with up to $10^5 \times 10^5$; and 3) a *Dual Attention* network (Sec. 3.3). It takes global and local features as inputs, brings interactions to them, and predicts gene expression.

Our contributions are summarized as follows: 1) a new **ISG** framework is proposed to predict gene expression from a histology slide image; 2) a new theoretical *Shannon Selection* module is proposed to filter out textureless image patches; 3) a new *Feature Extraction* network is proposed to extract discriminative patch features in an unsupervised manner; 4) a new *Dual Attention* network is proposed to calibrate patch features by injecting global contextual information to features and make better predictions; and 5) Our model outperforms

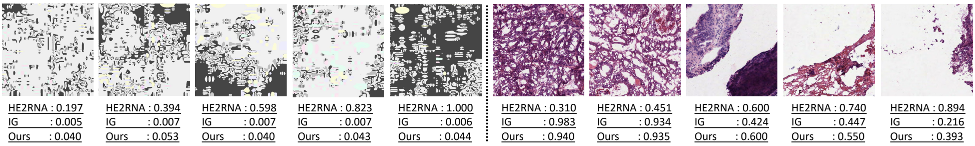


Figure 2: *Comparison of the proposed Shannon Selection module with HE2RNA [29] and the conventional image gradient (IG) based selection method.* ***Columns*** 1st-5th*: textureless patches. Note that HE2RNA fails to assign consistent scores for these patches and the scores are high. Both IG and our Shannon Selection module assign consistent low scores for these patches.* ***Columns*** 6th-10th*: patches with various degree of textures/features. Though both IG and our Shannon Selection module assign high scores for these patches, they prioritize texture/feature abundance differently (e.g.,* 8th *and* 9th *columns), resulting in different filtered patches and prediction accuracies (See comparison in Tab. 2).*

state-of-the-arts (SOTA) methods significantly.

2 Related Works

Computational Biomedical Domain. Deep learning demonstrates significant milestones in assisting disease diagnosis including cancer classification [8, 24], biomedical image segmentation [15, 22], tumor mutational burden prediction [3, 28], and gene expression prediction [4, 6, 29]. Gene expression prediction is the most essential and attractive task that would facilitate understanding and designing novel treatments [6]. Two sub-problems have been derived for two distinct gene expression profiling techniques [31, 35]. First, Schmauch *et al.* present a HE2RNA to model bulk RNA-Seq [35]. It targets quantifying gene expression for a whole histology sample which is up to $10^5 \times 10^5$ resolution (as a reference, this resolution is larger than the majority of remote sensing images [19]). They introduce a three-stage solution, including K-means and transfer learning, to extract image-level features. However, the model performance is promising to be further improved with a more task-specific design. Second, Dawood *et al.* [4], He *et al.* [6], and Zeng *et al.* [41] introduce NSL, STNet, and Hist2ST to measure spot-level gene expression of a histology slide image from a spatial transcriptomics (ST)-based [31] dataset. The ST technique is still under development, and there remains a wide audience to bulk RNA-Seq. Meanwhile, existing ST datasets lack diversities that typically contain tens of patients [6]. Thus, this paper follows HE2RNA [29] to model bulk RNA-Seq from histology slide images.

Representational Learning. There has been a huge effort from computer vision community on studying unsupervised feature learning [2, 14, 17, 21, 40]. With data augmentation crafted image views, a contrastive learning framework [2, 14, 17] usually learns a similarity-based representation from positive and negative matched view pairs. Alternatively, it enforces a unit cross-correlation matrix that is calculated from embeddings of two views of the same image [40]. However, these methods are hard to train because of requiring a large batch size and occasionally confront model collapses [21]. Instead, a StyleGAN [10, 11] has high accessibility, and it delivers a low dimension and versatile feature representation of a high-resolution image [27, 37, 38]. This is inevitable before establishing interactions of regions/patches among a histology slide image. In this paper, we investigate the use of StyleGAN for pre-learning gene expression features.

3 Methodology

Problem Formulation. Given a histology slide image $\mathbf{X} \in \mathbb{R}^{\mathsf{h}\times\mathsf{w}\times 3}$, we aim to predict its associated gene expression $\mathbf{Y} \in \mathbb{R}^{\mathsf{n}\times 1}$, where h, w, and n are height, width, and number of gene types, respectively. Our framework is given in Fig. 1.

We first tile $\mathbf{X}$ into patches at a coarse-resolution and build a patch set $\mathcal{X}^G = \{\mathbf{x}_i \in \mathbb{R}^{\mathsf{p}\times\mathsf{p}\times 3} \mid i \in 1, \cdots, \lfloor \frac{\mathsf{h}\times\mathsf{w}}{\mathsf{p}^2} \rfloor\}$, where p is the coarse patch size and $\lfloor\cdot\rfloor$ denotes the floor operator. $\mathcal{X}^G$ is further fed to a selector $\mathbf{S}(\cdot)$ to filter out textureless patches, resulting in a subset $\ddot{\mathcal{X}}^G = \mathbf{S}(\mathcal{X}^G)$ (Sec. 3.1). For each $\mathbf{x}_i \in \ddot{\mathcal{X}}^G$, we tile it into fine-grained patches at a fine-resolution, resulting in $\ddot{\mathcal{X}}_i^L = \{\mathbf{x}_{i,j} \in \mathbb{R}^{\mathsf{q}\times\mathsf{q}\times 3} \mid j \in 1, \cdots, \lfloor \frac{\mathsf{p}^2}{\mathsf{q}^2} \rfloor\}$, where q is the fine patch size and $\mathsf{q} < \mathsf{p}$. Collecting all fine patches yields a fine patch set $\ddot{\mathcal{X}}^L$.

Given $\ddot{\mathcal{X}}^G$ and $\ddot{\mathcal{X}}^L$, we separately use two feature extractors $\mathbf{E}^G(\cdot)$ and $\mathbf{E}^L(\cdot)$ to extract patch-wise low-dimension feature representation (Sec. 3.2). For each patch $\mathbf{x}_i \in \ddot{\mathcal{X}}^G$, we have $\mathbf{f}_i = \mathbf{E}^G(\mathbf{x}_i)$ and $\mathbf{f}_i \in \mathbb{R}^{\mathsf{d}\times 1}$, where d is the feature dimension. Collecting all coarse patch feature vectors yields a global feature set $\mathcal{F}^G = \{\mathbf{f}_i \mid i \in 1, \cdots, |\ddot{\mathcal{X}}^G|\}$. For each fine-grained patch set $\ddot{\mathcal{X}}_i^L \in \ddot{\mathcal{X}}^L$, it is corresponding to the patch $\mathbf{x}_i$. We extract patch-wise feature vectors for each fine-grained patch, arrange them according to their relative positions within $\mathbf{x}_i$ and obtain a feature map $\mathbf{R}_i = \mathbf{E}^L(\ddot{\mathcal{X}}_i^L)$, where $\mathbf{R}_i \in \mathbb{R}^{\frac{\mathsf{p}}{\mathsf{q}}\times\frac{\mathsf{p}}{\mathsf{q}}\times\mathsf{d}}$. Collecting all fine patch feature maps yields a local feature set $\mathcal{F}^L = \{\mathbf{R}_i \mid i \in 1, \cdots, |\ddot{\mathcal{X}}^G|\}$.

With $\mathcal{F}^G$ and $\mathcal{F}^L$, a dual attention network $\mathbf{M}(\cdot,\cdot)$ (Sec. 3.3) is proposed to fuse them and predicts the gene expression $\mathbf{Y} = \mathbf{M}(\mathcal{F}^G, \mathcal{F}^L)$.

3.1 Patch Selection Module

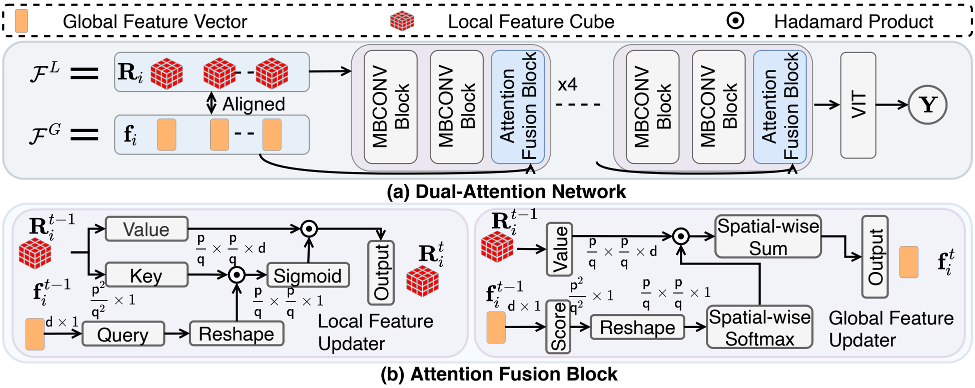
A histology slide image has large textureless/featureless regions. It is natural to select feature abundant patches by using edges, as edges are units to form features [30]. A number of edge detectors are available, e.g., Image Gradient (IG), Canny Edge Detector (CANNY), DexiNed (Dex) [26], Difference of Gaussian (DoG), and Laplacian of Gaussian (LoG). Instead of using edges, in this work, we give an alternative view from information theory to identify feature abundant patches. The comparison of our method and edge detection-based patch selectors are given in Sec. 4.1.

Shannon selection. For a patch $\mathbf{x}_i$, we use the Shannon information content [20] to quantify its texture/feature abundance level. The quantity is given by $h(\mathbf{x}_i) = \log_2 \frac{1}{\Pr(\mathbf{x}_i)}$, where $\Pr(\cdot)$ is the probability mass function of $\mathbf{x}_i$. Note, $h(\mathbf{x}_i)$ measures bit quantities of a patch. For a patch with poor features, $h(\mathbf{x}_i) \to 0$.

The key is to find a $\Pr(\cdot)$ that describes $\mathbf{x}_i$ well. Following [9, 25], we employ Solomonoff's universal prior $\Pr(\mathbf{x}_i) = \sum_{p:\mathcal{U}(p)=\mathbf{x}_i*} 2^{-\|p\|_0}$ as our patch distribution, where $\|\cdot\|_0$ is the length calculator, p is a program fed into a universal Turing machine $\mathcal{U}(\cdot)$, and $*$ denotes any possible suffix. This prior considers each feasible program p that derives '$\mathbf{x}_i*$', i.e., any string starts with a bit representation of $\mathbf{x}_i$, from $\mathcal{U}(\cdot)$. Afterwards, it sums over the negative exponent of the program length.

$$h(\mathbf{x}_i) = -\log \sum_{p:\mathcal{U}(p)=\mathbf{x}_i*} 2^{-\|p\|_0} \approx -\log 2^{-K(\mathbf{x}_i)} = K(\mathbf{x}_i), \tag{1}$$

where $K(\cdot)$ is the Kolmogorov complexity, and $K(\mathbf{x}_i)$ is the shortest program length for the input $\mathbf{x}_i$ and an excellent approximation of the Solomonoff's prior [9]. Eq. (1) suggests that a



(a) Dual-Attention Network

(b) Attention Fusion Block

Figure 3: *The architecture of our Dual Attention network.* ***(a)*** *For each $\mathbf{R}_i$ and $\mathbf{f}_i$ pair, we refine them with MBCONV [32] and attention fusion blocks, followed by a small ViT [5] for predicting the gene expression.* ***(b)*** *An attention fusion block consists of a local feature updater (Left) and a global feature updater (Right). A local feature updater uses the global feature vector $\mathbf{f}_i$ as the guidance, updating local features to emphasize features from regions of interest. A global feature updater uses the local feature $\mathbf{R}_i$ as the guidance, updating global features to reflect the evolving significance of local features.*

patch $\mathbf{x}_i$ deriving a large bit quantity tends to have abundant features. With a preset threshold, we select patches with abundant features.

3.2 Feature Extraction Network

For selected patches, we extract low-dimensional patch-wise features in this section. We use a style-based architecture for our feature extractor as it can capture versatile feature representations in an unsupervised manner [27, 37, 38]. Here, our extractor is from training a styleGAN-based autoencoder with image reconstruction as an auxiliary task.

Method. Let $\mathbf{D}(\cdot)$ be a styleGAN (aka. decoder) [10] and $\mathbf{E}(\cdot)$ be an associated extractor (aka. encoder) [38]. We use two different groups of them, $\mathbf{D}^G(\mathbf{E}^G(\cdot))$ and $\mathbf{D}^L(\mathbf{E}^L(\cdot))$, to learn salient features $\mathcal{F}^G$ and $\mathcal{F}^L$ from global and local patches, respectively. When extracting coarse-level features, $\mathbf{E}^G(\cdot)$ takes global patches as inputs. When extracting fine-grained features, $\mathbf{E}^L(\cdot)$ takes local patches as inputs. We empirically verify the versatility and effectiveness of our global and local feature extractors in Sec. 4.1.

Objectives. For brevity, we omit the superscript of $\mathbf{D}(\cdot)$ and $\mathbf{E}(\cdot)$, as the global and local feature extractors are trained in the same manner. To train them, we use the $\mathcal{L}_1$ loss, the AlexNet-based LPIPS loss $\mathcal{L}_{\text{LPIPS}}$ [13, 42], and the discriminator loss $\mathcal{L}_C$ (with a discriminator $\mathbf{C}(\cdot)$) [10], where $\mathcal{L}_1 = \|\mathbf{x} - \mathbf{D}(\mathbf{E}(\mathbf{x}))\|$, $\mathcal{L}_{\text{LPIPS}} = \|\phi(\mathbf{x}) - \phi(\mathbf{D}(\mathbf{E}(\mathbf{x})))\|_2$, and $\mathcal{L}_C = u\big(\mathbf{C}(\mathbf{x})\big) + u\big(-\mathbf{C}(\mathbf{D}(\mathbf{E}(\mathbf{x})))\big)$. $u(\cdot)$ is the Softplus function and $\phi(\cdot)$ is a pretrained LPIPS network.

The $\mathcal{L}_1$ loss and the LPIPS loss ensure our image reconstruction fidelity while affecting feature extraction quality [27, 38]. The discriminator loss is an indispensable objective function for style-based architecture. Our final training objective is given by

$$\mathcal{L}_{\text{total}} = \min_{\mathbf{D},\mathbf{E}} \max_{\mathbf{C}} \mathbb{E}_{\mathbf{x} \sim \mathbf{X}} \big[\mathcal{L}_1 + \mathcal{L}_{\text{LPIPS}} + \mathcal{L}_C\big]. \tag{2}$$

3.3 Dual Attention Network

With global features $\mathcal{F}^G = \{\mathbf{f}_i \mid i \in 1, \cdots, |\ddot{\mathcal{X}}^G|\}$ and local features $\mathcal{F}^L = \{\mathbf{R}_i \mid i \in 1, \cdots, |\ddot{\mathcal{X}}^G|\}$, we propose a *Dual Attention* network to adaptively calibrate model attention to regions of interest.

Method. The architecture of our *Dual Attention* network is given in Fig. 3. It has two modules connected in a sequence: 1) a lightweight MBCONV block sequence [32] interleaved with attention fusion blocks to jointly refine the local feature cube $\mathbf{R}_i$ and global feature $\mathbf{f}_i$; and 2) a small vision transformer [5] take average-pooled local feature cubes as input and predict the gene expression. Our attention fusion block has two components, the local feature updater and the global feature updater.

(a) Local Feature Updater. Let t be the layer index. Each feature vector in the local feature cube $\mathbf{R}_i^{t-1}$ represents a local patch, while having different priorities for predicting the gene expression. With the guidance from the global feature $\mathbf{f}_i^{t-1}$, we calibrate the local feature cube $\mathbf{R}_i^{t-1}$.

Specifically, we first project the global feature vector $\mathbf{f}_i^{t-1} \in \mathbb{R}^{\mathsf{d} \times 1}$ to obtain the query $\mathbf{Q}_i^t$ and the local feature cube $\mathbf{R}_i^{t-1} \in \mathbb{R}^{\frac{\mathsf{p}}{\mathsf{q}} \times \frac{\mathsf{p}}{\mathsf{q}} \times \mathsf{d}}$ to obtain the key $\mathbf{K}_i^t$ and the value $\mathbf{V}_i^t$.

$$\mathbf{Q}_i^t = \text{Reshape}(\mathbf{W}_q^t \mathbf{f}_i^{t-1}), \qquad \mathbf{K}_i^t = \mathbf{R}_i^{t-1} \mathbf{W}_k^t, \qquad \mathbf{V}_i^t = \mathbf{R}_i^{t-1} \mathbf{W}_v^t, \tag{3}$$

where $\mathbf{W}_q^t \in \mathbb{R}^{\frac{\mathsf{p}^2}{\mathsf{q}^2} \times \mathsf{d}}, \mathbf{W}_k^t \in \mathbb{R}^{\mathsf{d} \times \mathsf{d}}$, and $\mathbf{W}_v^t \in \mathbb{R}^{\mathsf{d} \times \mathsf{d}}$ are weight matrices.

We then compute the correlation of the query and key, followed by a Sigmoid activation function, to obtain a score map $\mathbf{A}_i^t$. Finally, the score map $\mathbf{A}_i^t$ is used to modulate the value $\mathbf{V}_i^t$ to obtain the updated local feature cube $\mathbf{R}_i^t$.

$$\mathbf{A}_i^t = \text{Sigmoid}\big(\mathbf{Q}_i^t \odot \mathbf{K}_i^t\big), \qquad \mathbf{Q}_i^t \in \mathbb{R}^{\frac{\mathsf{p}}{\mathsf{q}} \times \frac{\mathsf{p}}{\mathsf{q}} \times 1},\ \mathbf{K}_i^t \in \mathbb{R}^{\frac{\mathsf{p}}{\mathsf{q}} \times \frac{\mathsf{p}}{\mathsf{q}} \times \mathsf{d}}, \tag{4}$$

$$\mathbf{R}_i^t = \big(\mathbf{A}_i^t \odot \mathbf{V}_i^t\big) \mathbf{W}_r^t, \qquad \mathbf{A}_i^t,\ \mathbf{V}_i^t \in \mathbb{R}^{\frac{\mathsf{p}}{\mathsf{q}} \times \frac{\mathsf{p}}{\mathsf{q}} \times \mathsf{d}},\ \mathbf{W}_r^t \in \mathbb{R}^{\mathsf{d} \times \mathsf{d}}. \tag{5}$$

Here, $\mathbf{W}_r^t$ is an weight matrix, $\odot$ and $\text{Sigmoid}(\cdot)$ are the Hadamard product and the Sigmoid function. Unlike vanilla attention mechanism [34], we directly scale the value $\mathbf{V}_i^t$ by the score map $\mathbf{A}_i^t$, instead of costly aggregating it with matrix multiplications.

(b) Global Feature Updater. With the guidance from the local feature cube $\mathbf{R}_i^{t-1}$, we calibrate the global feature vector $\mathbf{f}_i^{t-1}$ to reflect the evolving significance of local features. Specifically, we first project the global feature vector $\mathbf{f}_i^{t-1}$ to the same spatial dimension with $\mathbf{R}_i^{t-1}$, and obtain a score matrix. We then normalize the score matrix spatially with a Softmax activation function to obtain a weight map $\mathbf{Z}_i^t$. Finally, we compute the Hadamard product between the weight map $\mathbf{Z}_i^t$ and projected local feature $\mathbf{P}_i^t$, followed by a sum-aggregation to obtain the updated global feature vector $\mathbf{f}_i^t$. Mathematically, we have

$$\mathbf{Z}_i^t = \text{Softmax}\big(\text{Reshape}(\mathbf{W}_z^t \mathbf{f}_i^{t-1})\big), \qquad \mathbf{W}_z^t \in \mathbb{R}^{\lfloor \frac{\mathsf{p}^2}{\mathsf{q}^2} \rfloor \times \mathsf{d}},\ \mathbf{f}_i^{t-1} \in \mathbb{R}^{\mathsf{d} \times 1}, \tag{6}$$

$$\mathbf{P}_i^t = \mathbf{R}_i^{t-1} \mathbf{W}_p^t, \qquad \mathbf{R}_i^{t-1} \in \mathbb{R}^{\frac{\mathsf{p}}{\mathsf{q}} \times \frac{\mathsf{p}}{\mathsf{q}} \times \mathsf{d}},\ \mathbf{W}_p^t \in \mathbb{R}^{\mathsf{d} \times \mathsf{d}}, \tag{7}$$

$$\mathbf{f}_i^t = \text{Sum}\big(\mathbf{Z}_i^t \odot \mathbf{P}_i^t\big) \mathbf{W}_f^t, \qquad \mathbf{Z}_i^t \in \mathbb{R}^{\frac{\mathsf{p}}{\mathsf{q}} \times \frac{\mathsf{p}}{\mathsf{q}} \times 1}, \mathbf{P}_i^t \in \mathbb{R}^{\frac{\mathsf{p}}{\mathsf{q}} \times \frac{\mathsf{p}}{\mathsf{q}} \times \mathsf{d}}, \mathbf{W}_f^t \in \mathbb{R}^{\mathsf{d} \times \mathsf{d}}, \tag{8}$$

where $\mathbf{W}_z^t$, $\mathbf{W}_p^t$, and $\mathbf{W}_f^t$ are weight matrices. $\text{Sum}(\cdot)$ denotes sum-aggregation along spatial dimensions, i.e., $\frac{\mathsf{p}}{\mathsf{q}} \times \frac{\mathsf{p}}{\mathsf{q}}$. $\text{Softmax}(\cdot)$ denotes the Softmax function that normalize scores along the same spatial dimensions.

Objectives. Similar to [29], we apply $\mathcal{L}_2$ loss to our *Dual Attention* network that establishes a mapping from global feature vectors $\mathcal{F}^G$ and local feature cubes $\mathcal{F}^L$ to gene expression $\mathbf{Y}$. We have $\mathcal{L}_2 = \| \mathbf{Y} - \mathbf{M}(\mathcal{F}^G, \mathcal{F}^L) \|^2$.

Table 1: *Gene expression predictions. We compare with SOTA methods using the standard Pearson Correlation Coefficient (PCC) metric. Our method consistently outperforms other methods for different gene types.*

Cancer Type	LIHC							COAD						PRAD			Avg.
Gene	CD3D	CD247	CD3E	CD3G	CD20	CD19	MK167	CD3D	CD247	CD3E	CD3G	CD20	CD19	TP63	KRT8	KRT18	
HE2RNA [29]	0.400	0.410	0.410	0.370	0.320	0.270	0.470	**0.430**	0.390	0.410	0.390	0.200	0.110	0.180	0.120	0.120	0.313
ViT-S [5]	0.193	0.252	0.256	0.279	0.258	0.187	0.189	0.260	0.291	0.312	0.300	0.314	0.280	0.065	0.153	0.151	0.234
ViT-MB[32]	0.337	0.388	0.378	0.345	0.360	0.361	0.464	0.351	0.379	0.382	0.370	0.379	0.383	0.207	0.154	0.165	0.338
CycleMLP[1]	0.343	0.364	0.396	0.374	0.353	0.352	0.348	0.320	0.347	0.378	0.377	0.372	0.378	0.203	0.140	0.187	0.327
MPViT[16]	0.365	0.358	0.377	0.350	0.379	0.356	0.491	0.311	0.361	0.371	0.352	0.374	0.413	0.205	0.142	0.199	0.338
ISG	**0.486**	**0.498**	**0.533**	**0.524**	**0.425**	**0.440**	**0.597**	0.415	**0.470**	**0.468**	**0.445**	**0.385**	**0.432**	**0.235**	**0.264**	**0.348**	**0.435**

4 Experiments

Dataset. We first evaluate our method on the popular dataset curated by Schmauch *et al.* [29], namely TCGA dataset. It has 6 different 'cancer + gene' type prediction tasks. In addition, to validate the generalization ability of our method, we directly apply our method to a clinic application, microsatellite instability (MSI) status prediction, on the three whole slide images (WSIs) dataset [12].

Methods for Comparison. We compare **ISG** framework with existing SOTA methods and possible alternatives (Tab. 1). 1) HE2RNA [29], which is the SOTA method in the gene expression prediction task. 2) ViT-S. We use ViT-S[1], which is a smaller version of ViT-B [5]. 3) ViT-MB, CycleMLP and MPViT. Both the ViT-S and ViT-B are unable to encode the local features directly, as the number of local features is quadratically increased with respect to the number of global features. Therefore, we use MBCONV blocks, CycleMLP [1], and MPViT [16] to downsample the local features for the ViT-S; and 4) Downsampled images. ViT architectures fail to converge if directly taking downsampled images as inputs.

Implementation details. Following [29], the target gene expression is log scale normalized, and we use the 5-fold cross-validation strategy as [29]. Each model is trained on assigned folds before tuning on each 'cancer + gene' type prediction task. Please refer to the supplementary material for more details of our model and the baselines.

4.1 Results

Comparisons with SOTA methods. We measure the Pearson Correlation Coefficient (PCC) [29] between model predictions and ground truth (GT) for gene types under different cancers. Our **ISG** achieves the best performance across all combinations of gene types and cancer types (Tab. 1). We have the following observations: 1) The low performance of ViT-S [5] indicates that only using global features ignores the fine-grained patch information; 2) After having the local feature, the performances of ViT-MB [32], CycleMLP [1], and MPViT [16] significantly improved in PCC. The improvement demonstrates the advantages of our extracted local patch features; and 3) Compared with the SOTA baseline HE2RNA [29], we achieve better performance, demonstrating the effectiveness of our ISG framework.

Comparison of Different Patch Selection Methods. Our **ISG** uses *Shannon Selection* module as the selector to find feature abundant patches. To demonstrate its effectiveness, we compare with commonly used patch selection methods, i.e., Image Gradient (IG), Canny Edge Detector (CANNY), DexiNed (Dex) [26], Difference of Gaussian (DoG), Laplacian of Gaussian (LoG), and Otsu algorithm (Otsu) [23]. They are denoted after the '-' symbol of **ISG**. The results are given in Tab. 2. We have the following observations: 1) our *Shannon*

[1] Experiments show that ViT-S achieves a better performance than the original ViT-B [5].

Table 2: *Comparison of using different patch selection methods.*

Cancer Type	LIHC							COAD						PRAD			Avg.
Gene	CD3D	CD247	CD3E	CD3G	CD20	CD19	MK167	CD3D	CD247	CD3E	CD3G	CD20	CD19	TP63	KRT8	KRT18	
ISG-IG	0.412	0.428	0.477	0.485	**0.432**	0.426	0.578	0.428	0.434	0.457	**0.449**	0.378	0.342	0.210	0.163	0.277	0.398
ISG-CANNY	0.473	0.445	0.511	0.484	0.373	0.396	0.583	0.400	0.436	0.436	0.438	0.340	0.341	0.190	0.249	0.318	0.401
ISG-Dex	0.436	0.437	0.503	0.493	0.428	0.433	0.576	0.386	0.423	0.442	0.427	0.338	0.362	0.189	0.245	0.315	0.402
ISG-DoG	0.468	0.431	0.490	0.457	0.411	0.414	0.591	0.382	0.431	0.422	0.421	0.350	0.351	0.209	**0.273**	0.343	0.403
ISG-LoG	0.338	0.293	0.316	0.337	0.305	0.323	0.354	0.181	0.235	0.216	0.207	0.243	0.199	0.136	0.155	0.202	0.253
ISG-Otsu	0.398	0.410	0.450	0.423	0.395	0.378	0.559	0.375	0.398	0.416	0.432	0.375	0.326	0.209	0.225	0.300	0.380
ISG	**0.486**	**0.498**	**0.533**	**0.524**	0.425	**0.440**	**0.597**	**0.415**	**0.470**	**0.468**	0.445	**0.385**	**0.432**	**0.235**	0.264	**0.348**	**0.435**

Table 3: *AUC of MSI status predictions.*

Dataset	**ISG**	HE2RNA [29]	MSIfromHE [12]
TCGA-CRC-DX	**0.86**	0.82	0.77
TCGA-CRC-KR	**0.87**	0.83	0.84
TCGA-STAD	0.78	0.76	**0.81**

Table 4: *Ablation study on Shannon Selection threshold.*

Threshold (bits)	8×10^5	1.6×10^6	2.4×10^6
Avg PCC $\uparrow$	0.386	**0.390**	0.373
Avg $\mathcal{L}_2 \downarrow$	0.024	**0.023**	0.025

Selection module achieves the best performance, with the averaged score at 0.435; and 2) with the same Otsu selector of HE2RNA [29], our method gets an average score of 0.380 and still outperforms HE2RNA (0.313), indicating the effectiveness of our full pipeline.

Extra Clinic Application. To validate the generalization ability of our method, we explore a direct clinic application: microsatellite instability (MSI) status prediction. We aim to distinguish MSI-High (MSI-H) from MSI-Stable (MSS). We use the standard area under the curve (AUC) [29] metric. The results on the datasets provided by [12] are given in Tab. 3. Our **ISG** outperforms HE2RNA [29] and achieves competitive performance compared with MSIfromHE [12] on each set of the WSIs dataset. Note that, our **ISG** model is directly applied on the WSIs dataset.

Efficiency. To compare the time efficiency of our model and HE2RNA, we estimate the inference time of 100 randomly sampled slide images for **ISG** and HE2RNA [29]. For fair comparisons, we use the same GPU and slide image reading package `slideio`. On average, the inference on each slide image takes 246 seconds by using our methods, while HE2RNA takes 302 seconds. HE2RNA squanders computations on extracting features from textureless patches, due to the Otsu selector failing to effectively filter these patches (see Sec. 1). Furthermore, we verify the generalizability ability of our proposed *Shannon Selection* Module. We apply the *Shannon Selection* module in the HE2RNA for selecting patches with abundant features. This simple replacement increases the average performance of 'HE2RNA-Shannon' by 7.5%.

4.2 Discussion

All experiments in this section are done by using a single model for all of the 6 prediction tasks in the TCGA dataset. Please refer to the supplementary material for more experiments and analysis.

Shannon Selection Threshold. We study the the selection threshold of the *Shannon Selection* module (Tab. 4). We achieve the best PCC with the threshold at 1.6×10^6 bits by following the mathematical implication. Suppose the patch texture abundance distribution follows a normal distribution $\mathcal{N}(\cdot \mid \mu, \sigma^2)$, where μ and σ^2 are mean and variance. Our filter roughly keeps patches residing in the area of being at least one positive standard deviation from the mean, i.e., those patches with abundance score bigger than $\mathcal{N}(\mu + \sigma^2 \mid \mu, \sigma^2)$. Semantically, the patch is known to be texture abundant. A smaller selection threshold results in massive textureless patches that imposes interference signals to model inference.

Table 5: *Representation effectiveness evaluations. '-' denotes the result is unavailable.*

Model	ResNet50 (Pretrained)	*Feature Extraction* network (Pretrained)	ResNet50 (Finetuned)
TCGA dataset [29]	0.123	**0.223**	-
STNet dataset [6]	0.036	**0.178**	0.162

Table 6: *Ablation Study on Dual Attention network.*

Model	**ISG**-6-2	**ISG**-8-2	**ISG**-10-2	**ISG**-12-2	**ISG**-14-2	**ISG**-10-1	**ISG**-10-2	**ISG**-10-3
Avg PCC ↑	0.3723	0.3750	**0.3908**	0.3601	0.3263	0.2488	**0.3908**	0.3628
Avg $\mathcal{L}_2$ ↓	**0.0229**	0.00231	0.0234	0.0233	0.0239	0.0262	0.0234	0.0235

Reversely, a larger selection threshold excludes patches with moderate texture abundance that restricts the model from perceiving comprehensive input image features. As shown in Tab. 4, our **ISG** is robust with the varying Shannon selection threshold. Moreover, even with a sub-optimal selection threshold and using the single model for all 6 prediction tasks, our framework outperforms HE2RNA in PCC (Tab. 1), i.e., 0.373 vs. 0.313.

Feature Effectiveness. To verify the quality of extracted features by our *Feature Extraction* network, we train a simple predictor, i.e., a two-layer perceptron with a ReLU activation. With this predictor, our *Feature Extraction* network is compared to a pretrained ResNet50 (suggested by HE2RNA [29]) in Tab. 5. We finetune a ResNet50 as a reference whenever possible. For the TCGA dataset [29], all patch representations are pooled to input the predictor. Using our proposed *Feature Extraction* network, the predictor demonstrates a stronger PCC than the pretrained ResNet50 representations. Note that we do not present the results of a finetuned ResNet50 because of GPU memory constraints. Second, we explore the STNet dataset [6]. It is a small-scale dataset that contains spot-level (149 × 149 pixels) gene expression annotations. We use the target gene expression types selected by [6]. Our proposed representations consistently beat both the pretrained and finetuned ResNet50 representations for the gene expression prediction task. Results demonstrate the robustness and expressiveness of our proposed *Feature Extraction* network across slide image-based datasets.

Architectures. We ablate the number of MBCONV blocks and frequency of using our proposed attention fusion blocks in our *Dual Attention* network (Tab. 6). Their configurations are denoted after '**ISG**' and separated by the '-' symbol. We measure their performance with PCC and $\mathcal{L}_2$. Our observations are: 1) **ISG**-10-2 tends to be an optimal setup that balances interactions between local features and global features to obtain the best PCC. However, **ISG**-6-2 leads to the best $\mathcal{L}_2$. Our task emphasizes the relative changes in gene expression, thus we bias on the PCC measurement and recommend **ISG**-10-2 as our final architecture; and 2) There is no strict correlation between PCC and $\mathcal{L}_2$. The former counts an integral correlation between the prediction and the GT across all samples. The latter calculates a sample-wise deviation between the predictions and the GT. Thus, they behave differently in measuring model performance.

5 Conclusion

In this paper, we have proposed an **ISG** framework to predict gene expression from histology slide images. Our key idea is to establish spatial interactions among sparsely and non-uniformly distributed feature patches of the input image for the prediction. To do so, we select the patches tiled at two distinct resolutions with abundant features by our *Shannon Selection* module. Then, the patches are embedded into low-dimension representations by our

Feature Extraction network trained in an unsupervised manner. Finally, we design a *Dual Attention* network to refine the extracted features, to calibrate network attention on the regions of interest for gene prediction. Extensive experiments have validated the effectiveness, efficiency, and generalization ability of our method. We hope the proposed **ISG** framework can facilitate disease diagnosis and treatment development.

Acknowledgment. Liyuan Pan's work was supported in part by the Beijing Institute of Technology Research Fund Program for Young Scholars.

References

[1] Shoufa Chen, Enze Xie, Chongjian Ge, Ding Liang, and Ping Luo. Cyclemlp: A mlp-like architecture for dense prediction. *CoRR*, abs/2107.10224, 2021. URL https://arxiv.org/abs/2107.10224.

[2] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 15750–15758. Computer Vision Foundation / IEEE, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Chen_Exploring_Simple_Siamese_Representation_Learning_CVPR_2021_paper.html.

[3] Nicolas Coudray, Paolo Ocampo, Theodore Sakellaropoulos, Navneet Narula, Matija Snuderl, David Fenyö, Andre Moreira, Narges Razavian, and Aristotelis Tsirigos. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature Medicine*, 24, 10 2018. doi: 10.1038/s41591-018-0177-5.

[4] Muhammad Dawood, Kim Branson, Nasir Rajpoot, and Fayyaz ul Amir Afsar Minhas. All you need is color: Image based spatial gene expression prediction using neural stain learning. 08 2021.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

[6] Bryan He, Ludvig Bergenstråhle, Linnea Stenbeck, Abubakar Abid, Alma Andersson, Ake Borg, Jonas Maaskola, Joakim Lundeberg, and James Zou. Integrating spatial gene expression and breast tumour morphology via deep learning. *Nature Biomedical Engineering*, 4:1–8, 08 2020. doi: 10.1038/s41551-020-0578-x.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90.

[8] Le Hou, Dimitris Samaras, Tahsin M. Kurç, Yi Gao, James E. Davis, and Joel H. Saltz. Patch-based convolutional neural network for whole slide tissue image classification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2424–2433. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.266. URL https://doi.org/10.1109/CVPR.2016.266.

[9] Marcus Hutter. Universal artificial intelligence: Sequential decisions based on algorithmic probability. 04 2012. doi: 10.1007/b138233.

[10] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 8107–8116. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.00813. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Karras_Analyzing_and_Improving_the_Image_Quality_of_StyleGAN_CVPR_2020_paper.html.

[11] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *CoRR*, abs/2106.12423, 2021. URL https://arxiv.org/abs/2106.12423.

[12] Jakob Kather, Alexander Pearson, Niels Halama, Dirk Jäger, Jeremias Krause, Sven Loosen, Alexander Marx, Peter Boor, Frank Tacke, Ulf Neumann, Heike Grabsch, Takaki Yoshikawa, Hermann Brenner, Jenny Chang-Claude, Michael Hoffmeister, Christian Trautwein, and Tom Luedde. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nature Medicine*, 25, 07 2019. doi: 10.1038/s41591-019-0462-y.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, 2017. doi: 10.1145/3065386. URL http://doi.acm.org/10.1145/3065386.

[14] Nikolina Kubiak, Armin Mustafa, Graeme Phillipson, Stephen Jolly, and Simon Hadfield. SILT: self-supervised lighting transfer using implicit image decomposition. *CoRR*, abs/2110.12914, 2021. URL https://arxiv.org/abs/2110.12914.

[15] Victor Kulikov and Victor S. Lempitsky. Instance segmentation of biological images using harmonic embeddings. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 3842–3850. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.00390. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Kulikov_Instance_Segmentation_of_Biological_Images_Using_Harmonic_Embeddings_CVPR_2020_paper.html.

[16] Youngwan Lee, Jonghee Kim, Jeffrey Willette, and Sung Ju Hwang. Mpvit: Multipath vision transformer for dense prediction. *CoRR*, abs/2112.11010, 2021. URL https://arxiv.org/abs/2112.11010.

[17] Yang Liu, Alexandros Neophytou, Sunando Sengupta, and Eric Sommerlade. Relighting images in the wild with a self-supervised siamese auto-encoder. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pages 32–40. IEEE, 2021. doi: 10.1109/WACV48630.2021.00008. URL https://doi.org/10.1109/WACV48630.2021.00008.

[18] Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with warm restarts. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=Skq89Scxx.

[19] Lei Ma, Yu Liu, Xueliang Zhang, Yuanxin Ye, Gaofei Yin, and Brian Johnson. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 152:166–177, 04 2019. doi: 10.1016/j.isprsjprs.2019.04.015.

[20] David J. C. MacKay. *Information theory, inference, and learning algorithms*. Cambridge University Press, 2003. ISBN 978-0-521-64298-9.

[21] Divyam Madaan, Jaehong Yoon, Yuanchun Li, Yunxin Liu, and Sung Ju Hwang. Rethinking the representational continuity: Towards unsupervised continual learning. *CoRR*, abs/2110.06976, 2021. URL https://arxiv.org/abs/2110.06976.

[22] Yanda Meng, Hongrun Zhang, Dongxu Gao, Yitian Zhao, Xiaoyun Yang, Xuesheng Qian, Xiaowei Huang, and Yalin Zheng. BI-GCN: boundary-aware input-dependent graph convolution network for biomedical image segmentation. *CoRR*, abs/2110.14775, 2021. URL https://arxiv.org/abs/2110.14775.

[23] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *Systems, Man and Cybernetics, IEEE Transactions on*, 9:62–66, 01 1979.

[24] Fábio Perez, Sandra Avila, and Eduardo Valle. Solo or ensemble? choosing a CNN architecture for melanoma classification. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 2775–2783. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPRW.2019.00336. URL http://openaccess.thecvf.com/content_CVPRW_2019/html/ISIC/Perez_Solo_or_Ensemble_Choosing_a_CNN_Architecture_for_Melanoma_Classification_CVPRW_2019_paper.html.

[25] Jan Poland and Marcus Hutter. Asymptotics of discrete mdl for online prediction. *Information Theory, IEEE Transactions on*, 51:3780 – 3795, 12 2005. doi: 10.1109/TIT.2005.856956.

[26] Xavier Soria Poma, Ángel D. Sappa, Patricio Humanante, and Arash Akbarinia. Dense extreme inception network for edge detection. *CoRR*, abs/2112.02250, 2021. URL https://arxiv.org/abs/2112.02250.

[27] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: A stylegan encoder for image-to-image translation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021,*

virtual, June 19-25, 2021, pages 2287–2296. Computer Vision Foundation / IEEE, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Richardson_Encoding_in_Style_A_StyleGAN_Encoder_for_Image-to-Image_Translation_CVPR_2021_paper.html.

[28] Andrew Schaumberg, Mark Rubin, and Thomas Fuchs. H&e-stained whole slide image deep learning predicts spop mutation state in prostate cancer. 10 2018. doi: 10.1101/064279.

[29] Benoit Schmauch, Alberto Romagnoni, Elodie Pronier, Charlie Saillard, Pascale Maillé, Julien Calderaro, Aurélie Kamoun, Meriem Sefta, Sylvain Toldo, Mikhail Zaslavskiy, Thomas Clozel, Matahi Moarii, Pierre Courtiol, and Gilles Wainrib. A deep learning model to predict rna-seq expression of tumours from whole slide images. *Nature Communications*, 11, 08 2020. doi: 10.1038/s41467-020-17678-4.

[30] E.U. Scott. Digital image processing and analysis: human and computer vision applications with cviptools. *Digital Image Processing and Analysis: Human and Computer Vision Applications with CVIP Tools*, 01 2011.

[31] Patrik Ståhl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, Jose Fernandez Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub Westholm, Mikael Huss, Annelie Mollbrink, Sten Linnarsson, Simone Codeluppi, Åke Borg, Fredrik Pontén, Paul Costea, Pelin Akan Sahlén, Jan Mulder, Olaf Bergmann, and Jonas Frisén. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353:78–82, 07 2016. doi: 10.1126/science.aaf2403.

[32] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 2019. URL http://proceedings.mlr.press/v97/tan19a.html.

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. 06 2017.

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[35] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: A revolutionary tool for transcriptomics. *Nature reviews. Genetics*, 10:57–63, 12 2008. doi: 10.1038/nrg2484.

[36] B. White and Frank Rosenblatt. Principles of neurodynamics: Perceptrons and the theory of brain mechanisms. *The American Journal of Psychology*, 76:705, 12 1963. doi: 10.2307/1419730.

[37] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 12863–12872. Computer Vision Foundation / IEEE, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Wu_StyleSpace_Analysis_Disentangled_Controls_for_StyleGAN_Image_Generation_CVPR_2021_paper.html.

[38] Yan Yang, Md. Zakir Hossain, Tom Gedeon, and Shafin Rahman. S2FGAN: semantically aware interactive sketch-to-face translation. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*, pages 3162–3171. IEEE, 2022. doi: 10.1109/WACV51458.2022.00322. URL https://doi.org/10.1109/WACV51458.2022.00322.

[39] Xiaoyu Yue, Shuyang Sun, Zhanghui Kuang, Meng Wei, Philip Torr, Wayne Zhang, and Dahua Lin. Vision transformer with progressive sampling. pages 377–386, 10 2021. doi: 10.1109/ICCV48922.2021.00044.

[40] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12310–12320. PMLR, 2021. URL http://proceedings.mlr.press/v139/zbontar21a.html.

[41] Yuansong Zeng, Zhuoyi Wei, Weijiang Yu, Rui Yin, Yuchen Yuan, Bingling Li, Zhonghui Tang, Yutong Lu, and Yuedong Yang. Spatial transcriptomics prediction from histology jointly through transformer and graph neural networks. *Briefings in Bioinformatics*, 07 2022. doi: 10.1093/bib/bbac297.

[42] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 586–595. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00068. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_The_Unreasonable_Effectiveness_CVPR_2018_paper.html.