Supplemental Materials for "ISG: I can See Your Gene Expression"

In this supplementary material, we show implementation details (Appx. A) for our frameworks and alternative baselines, extra experiments (Appx. B), and histology slide image samples (Appx. C).

A Implementation Details

We tile local and global patch sizes at resolutions of 256 and 2048, and patches have 10% overlapping ratio for height and width (see Tab. 7 for verification of this setup). For *Shannon Selection* module, our selection threshold is 1.6×10^6 bits. We approximate Kolmogorov complexity with a `Pillow PNG` compressor. For the *Feature Extraction* network, the latent dimension of the local feature extractor $\mathbf{E}^L(\cdot)$ and the global feature extractor $\mathbf{E}^G(\cdot)$ is 256 and 512, respectively. Comparing to $\mathbf{E}^L(\cdot)$, $\mathbf{E}^G(\cdot)$ has a extra four times downsampling layer at the beginning. These *Feature Extraction* networks follow optimization schemes and architecture in [38]. For *Dual Attention* network, we train our model from scratch under Adam optimizer for 200 epochs on assigned folds, and it then is tuned for 100 epochs on each 'cancer + gene' type prediction task. We have a batch size of 32. The learning rate is set to 5×10^{-4} and is decayed with a cosine annealing learning rate scheduler [18]. We use 10 MBCONV blocks [32] and insert an 8-head attention fusion block with residual connections after every two MBCONV blocks, where the attention fusion block configurations are automatically determined (Sec. 3.3). By the naming convention from [32], the MBCONV blocks have:

Expanding ratio	Out channel	Number	stride	SELayer
1	256	2	1	1
1	256	2	2	1
1	304	2	1	1
1	384	2	1	1
1	512	2	1	1

In the following, we apply a global average pooling to each local feature cube and feed the outputs into a smaller version of ViT-B [5], i.e., ViT-S. This ViT-S sets the patch feature dimension, depth, feed-forward dimension, and the head number to 512, 6, 1024, and 8, respectively.

We apply the same optimization settings for the baseline methods which are: 1) ViT-S. It follows the above configurations. 2) ViT-MB [32]. It combines the ViT-S and MBCONV blocks above. However, as solely MBCONV blocks cannot have interactions between the local features and global features. We linearly project concatenation of the global features and the pooled local feature cube outputs from the MBCONV blocks before feeding into the ViT-S. 3) CycleMLP [1]. It is similar to the ViT-MB. The CycleMLP layers are configured with

Transition	Layer	MLP ratio	embed dim
True	2	4	256
True	2	4	384
True	4	4	384
True	2	4	512

where patch size for the CycleMLP is one, and the remaining settings follow 'cycle_s' [1]. Please refer to [1] for more details. 4) MPViT. It replaces the MBCONV blocks in ViT-MB. Following [16], the configurations are summarised as

Path	Layer	MLP ratio	embed dim	Head
2	4	2	256	8
2	6	2	512	8

B Extra Experiments

We experimentally search for an optimal overlapping ratio between patches (Tab. 7). We select the ratio based on the average performance of **ISG** across different 'gene + cancer' type tasks. The 10% overlapping ratio leads to the most performing version of **ISG**.

Cancer Type	LIHC							COAD						PRAD			Avg.
Gene	CD3D	CD247	CD3E	CD3G	CD20	CD19	MK167	CD3D	CD247	CD3E	CD3G	CD20	CD19	TP63	KRT8	KRT18	
0%	0.426	0.447	0.527	0.472	0.413	0.392	0.559	0.432	0.398	0.428	0.421	0.386	0.450	0.216	0.276	0.384	0.414
10%	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**
20 %	0.476	0.464	0.513	0.504	0.421	0.422	0.594	0.438	0.415	0.434	0.416	0.383	0.422	0.232	0.258	0.355	0.422
30 %	0.461	0.423	0.498	0.483	0.431	0.437	0.577	0.415	0.482	0.470	0.447	0.391	0.373	0.262	0.219	0.325	0.418

Table 7: The performance change with respect to overlapping ratios of patches.

C Histology Slide Images

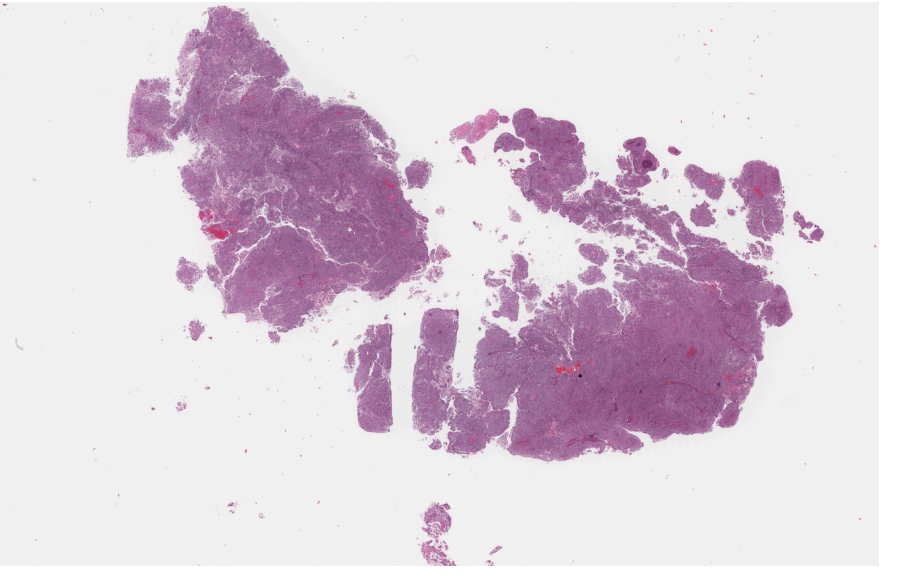

Figure 4: A histology slide image sample.

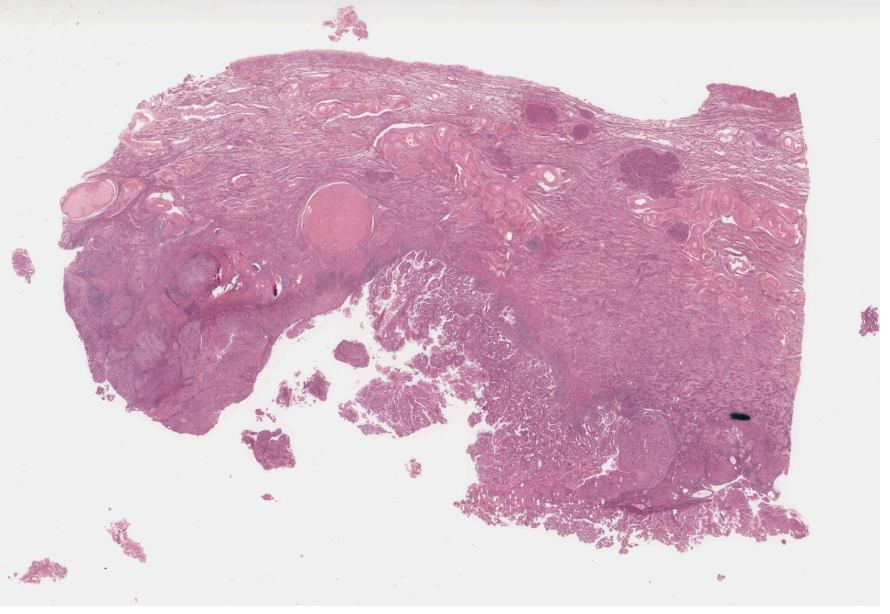

Figure 5: A histology slide image sample.

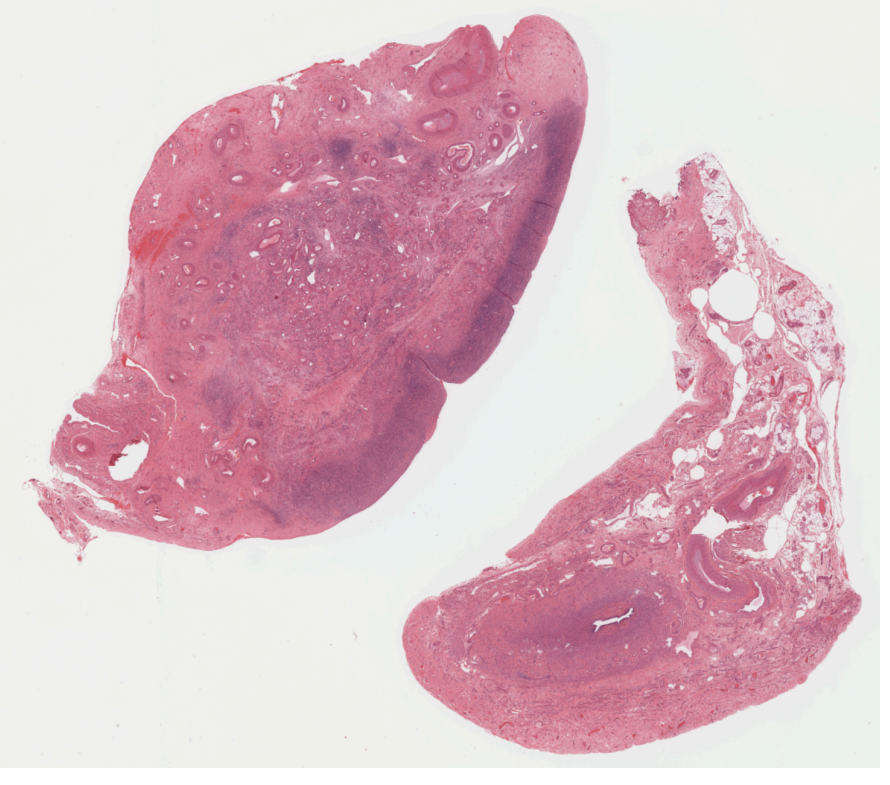

Figure 6: A histology slide image sample.