Supplementary Material

A. Multi-Head Self-Attention

We briefly introduce self-attention then extend it to the multi-head version. Recall that $\{\mathbf{x}_l^{(i)}\}_{i=1}^{M} \in \mathbb{R}^{M \times d_l}$ denotes the sequence of visual token extracted from a $L \times L$ patch, where M is the sequence length and d_l is the embedding dimension extracted for l-sized tokens. For simplicity, we ignore the the subscript l and lower the superscript i to the subscript. The sequence of visual token is re-written as $S = \{\mathbf{x}_i\}_{i=1}^{M} \in \mathbb{R}^{M \times d_l}$

To perform self-attention, we add three matrices $\boldsymbol{W}_q$,$\boldsymbol{W}_k$ and $\boldsymbol{W}_v \in \mathbb{R}^{d_l \times d_l}$ used to compute the relationship between the token itself $\mathbf{x}_i$ and other tokens $\mathbf{x}_j$ in S. Specifically,

$$\boldsymbol{q}_i = \boldsymbol{W}_q \mathbf{x}_i, \boldsymbol{k}_i = \boldsymbol{W}_k \mathbf{x}_i, \boldsymbol{v}_i = \boldsymbol{W}_v \mathbf{x}_i$$

$$w_{ij} = \text{softmax}(\frac{\boldsymbol{q}_i^\intercal \boldsymbol{k}_j}{\sqrt{d_l}})$$

$$\boldsymbol{o}_i = \sum_j w_{ij} \boldsymbol{v}_j,$$

where $\boldsymbol{q}_i$, $\boldsymbol{k}_i$ and $\boldsymbol{v}_i$ are query, key and value vector of input token $\mathbf{x}_i$. Intuitively, each visual token $\mathbf{x}_i$ computes a similarity score with other tokens and use the normalized similarity score to perform weighted-sum of value vectors of other tokens. The query, key and value vector space are captured by $\boldsymbol{W}_q$,$\boldsymbol{W}_k$ and $\boldsymbol{W}_v$ learned from data. To increase the feature expressiveness, we increase the number of matrices from 1 to h which leads to a set of matrices $\{\boldsymbol{W}_q^{h'}, \boldsymbol{W}_k^{h'}, \boldsymbol{W}_v^{h'} | h' = 1, 2, 3, \ldots, h\}$ for the input sequence. The self-attention with $h > 1$ is called multi-head self-attention (MHSA), which we use not only as the building block in permutation-equivariant aggregation of visual tokens, but also demonstrate to learn representations of morphological phenotypes, as seen in Figure 1.

B. Additional Implementation Details

Efficient Inference and Batching in HIPT. At inference time, we feed in $\mathbf{x}_{\text{WSI}} \in \mathbb{R}^{M \times 256 \times 256 \times [3 \times 16 \times 16]}$ with a slide-level batch size $B_{\text{WSI}} = 1$ into HIPT, in which the respective dimensions correspond to: length of $\mathbf{x}_{4096}$ regions in $\mathbf{x}_{\text{WSI}}$, length of $\mathbf{x}_{256}$ patches in $\mathbf{x}_{4096}$, length of $\mathbf{x}_{16}$ cells in $\mathbf{x}_{256}$, and the remaining dimensions being the $\mathbf{x}_{16}$ shape. Without pre-extracting any tokens, inference can be performed in first defining a DataLoader class over a $M \times 3 \times 4096 \times 4096$ view of $\mathbf{x}_{\text{WSI}}$ with $B_{4096} = 1$ to iterate over single $\mathbf{x}_{4096}$ regions, followed by using Einsum operations to unroll each region via Einsum : $\mathbb{R}^{1 \times 3 \times 4096 \times 4096} \rightarrow \mathbb{R}^{256 \times 3 \times 256 \times 256}$. Lastly, another DataLoader is defined over this view using the first dimension as $B_{256} = 256$, which then leads to the bottom-up aggregation strategy starting with the ViT_{256}-16 forward pass.

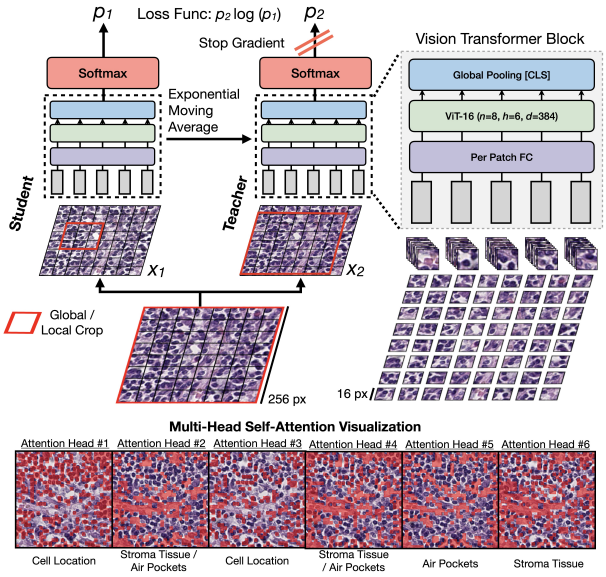

Figure 1. **ViT_{256}-16 DINO Pretraining**. Self-supervision using knowledge distillation in DINO to pretrain ViT_{256}-16 on histology image patches [2]. A student network $\phi_{s_{256}}$ is trained to match the probability distribution of a Siamese teacher network $\phi_{t_{256}}$ using a cross-entropy loss, with ϕ parameterized using a ViT_{256}-16 model, and local and global crops applied as data augmentation. Interpretability of multi-head attention weights reveal that DINO learns distinct morphological phenotypes. In "red" are high-attention visual tokens with attention weights greater than 0.5.

Pretraining Dataset Curation:. As noted in the main paper, we pretrain ViT_{256}-16 and ViT_{4096}-256 in different stages using 10,687 FFPE (formalin-fixed, paraffin-embedded) H&E-stained diagnostic slides from 33 cancer types in the The Genome Cancer Atlas (TCGA), and extracted 408,218 $\mathbf{x}_{4096}$ regions at an $20\times$ objective ($M \approx 38$ regions per slide) for pretraining ViT_{4096}-256, with a total of 104 Million $\mathbf{x}_{256}$ patches for pretraining ViT_{256}-16 [5]. To curate these regions, we used the Tissue Image Analysis (TIA) Toolbox to patch slides into non-overlapping, tissue-containing regions, with additional quality control performed to limit regions that contained predominantly background slide information (*e.g.* - white space). A limitation of this study is that since our method requires $[4096 \times 4096]$ patching, not all slides from the TCGA were used for pretraining and weakly-supervised evaluation. A full table showing number of WSIs and regions used is shown on a per-cancer basis in Table 1. As efficient stor-

age, $\mathbf{x}_{4096}$ images for each slide were stored using TAR archives using the WebDataset API, which we plan to make available in a public release.

Attention Visualization:. To create attention heatmaps, we follow the work of Caron *et al*. using the attention map for each head at the last ViT stage, and linearly interpolate the attention map such that the attention score each token is $[16 \times 16]$ for ViT_{256}-16 and $[256 \times 256]$ for ViT_{4096}-256. No Gaussian blurring or other smoothing operations were performed. To obtain more granular maps for ViT_{4096}-256, we computed attention scores for each patch using an overlapping stride length of 64. To create hierarchical attention maps, we factorized the attention weight distribution within each weighted $\mathbf{x}_{256}$ patch from ViT_{4096}-256 with the attention weight distribution of ViT_{256}-16. These hierarchical attention maps can be interpreted as: *For weighted* $\boldsymbol{x}_{256}$ *patches localized by* ViT_{4096}-256 *by head X, what are the important* $\boldsymbol{x}_{16}$ *localized within that patch by head Y in* ViT_{256}-16*?* This allows the creation of certain attention maps that localize: 1) $[16 \times 16]$ tumor cells within 256×256 stromal regions, or 2) $[16 \times 16]$ tumor cells within 256×256 poorly-differentiated glands / larger tumor structures. With $h = 6$ for both ViT_{256}-16 and ViT_{4096}-256, a total of 36 hierarchical attention maps can be created. From pathologist inspection, 2-3 heads in each VIT model localized unique morphological phenotypes.

Dataset	# Slides	# Regions	Size (in GB)
ACC	223	12254	230.9
BLCA	454	21381	389.1
BRCA	1038	30248	521.1
CESC	271	8846	152.8
CHOL	38	2214	42.2
COADREAD	588	14705	249.3
ESCA	145	5820	103.9
GBMLGG	1541	54158	932.5
HNSC	451	17204	288.6
KIR(C/P/CH)	918	38019	705.5
LIHC	375	19358	369.1
LUADLUSC	1008	43487	757.3
LYM	43	1590	28.3
MESO	81	2521	43.9
OV	107	5222	95.1
PAAD	204	7377	133.2
PRAD	421	17171	301.9
SARC	567	26974	503.7
SKCM	456	19415	351.7
STAD	371	14664	253.0
TGCT	233	8389	154.8
THCA	516	26611	418.5
UCEC	561	34494	628.0
UVM	77	1657	26.7
Total	10687	433779	7.7 TB

Table 1. **TCGA Pan-Cancer Datasheet.** Total number of slides, 4096×4096 image regions, and their storage size.

C. Additional Quantitative Experiments

C.1. Variations in HIPT Architecture

Model Descriptions. We performed ablation experiments assessing the most impactful components of the HIPT architecture, as demonstrated in the slide-level classification results in Table 2. Specifically, we inspected variations in the HIPT architecture:

- $\mathbf{ViT_{256}\text{-}16_{PF}}$, **AP-256, AP-4096**: This variation uses no ViT components as permutation-equivariant aggregation hidden layers at the patch- and region-level. Features are still pre-extracted using a ViT_{256}-16 (denoted with "P" for pretrained, "F" for frozen), however, only attention pooling (AP) is performed across the different stages.

- $\mathbf{ViT_{256}\text{-}16_{PF}}$, $\mathbf{ViT_{4096}\text{-}256}$, $\mathbf{ViT_{WSI}\text{-}4096}$: This variation uses ViT_{4096}-256 and ViT_{WSI}-4096 for patch- and region-level aggregation, however, these aggregation layers are trained from scratch.

- $\mathbf{ViT_{256}\text{-}16_{PF}}$, $\mathbf{ViT_{4096}\text{-}256_{P}}$, $\mathbf{ViT_{WSI}\text{-}4096}$: This variation uses ViT_{4096}-256 and ViT_{WSI}-4096 for patch- and region-level aggregation, withViT_{4096}-256 additionally pretrained using Stage 2 Hierarchical Pretraining.

- ViT_{256}-16_{PF}, ViT_{4096}-256_{PF}, ViT_{WSI}-4096: This variation uses ViT_{4096}-256 and ViT_{WSI}-4096 for patch- and region-level aggregation, with ViT_{4096}-256 additionally pretrained and frozen. Only parameters for ViT_{WSI}-4096 are finetuned.

Pretraining and Freezing Prevents Overfitting. Across slide-level classification tasks, we observe that pretraining and freezing ViT_{4096}-256 in HIPT is an important component an achieving strong performance. Without freezing, training the parameters for ViT_{4096}-256 results in the total number of trainable parameters to be 3388996. Though small for a Transformer model, training and finetuning on WSI datasets with less than 1000 data points may easily result in overfitting, as demonstrated in Table 2. In particular, we observe that without freezing, performance drops from 0.923 to 0.652 and 0.952 to 0.820 on NSCLC subtyping with 25% and 100% training data finetuning respectively. We also compare HIPT to a variation that does not

Architecture	# Params	BRCA Subtyping		NSCLC Subtyping		RCC Subtyping	
		25% Training	100% Training	25% Training	100% Training	25% Training	100% Training
ViT-16$_{PF}$, AP-256, AP-4096	494597	0.784 ± 0.061	0.837 ± 0.062	0.835 ± 0.050	0.928 ± 0.023	0.955 ± 0.016	0.965 ± 0.013
ViT-16$_{PF}$, ViT-256, ViT-4096	3388996	0.758 ± 0.076	0.823 ± 0.071	0.695 ± 0.069	0.786 ± 0.096	0.928 ± 0.038	0.956 ± 0.016
ViT-16$_{PF}$, ViT-256$_{P}$, ViT-4096	3388996	0.762 ± 0.089	0.827 ± 0.069	0.652 ± 0.076	0.820 ± 0.047	0.935 ± 0.022	0.956 ± 0.013
ViT-16$_{PF}$, ViT-256$_{PF}$	505204	**0.821 ± 0.069**	**0.874 ± 0.060**	**0.923 ± 0.020**	**0.952 ± 0.021**	**0.974 ± 0.012**	**0.980 ± 0.013**
ResNet-50$_{B3, IN}$, GMP	-	0.638 ± 0.089	0.667 ± 0.070	0.696 ± 0.055	0.794 ± 0.035	0.862 ± 0.030	0.951 ± 0.016
ViT-16$_{PF}$, GMP	-	0.605 ± 0.092	0.725 ± 0.083	0.622 ± 0.067	0.742 ± 0.045	0.848 ± 0.032	0.899 ± 0.027
ViT-16$_{PF}$, ViT-256$_{PF}$, GMP	-	**0.682 ± 0.055**	**0.775 ± 0.042**	**0.773 ± 0.048**	**0.889 ± 0.027**	**0.916 ± 0.022**	**0.974 ± 0.016**

Table 2. **Slide-Level Classification. Top Row.** Ablation study assessing impact of Transformer attention, pretraining, and parameter freezing in the HIPT architecture. **Bottom Row.** Ablation study assessing K-Nearest Neighbors (K-NN) performance using the average pre-extracted embeddings with different pretrained embedding types. **Abbreviations.** "P" = Pretrained. "F" = Frozen. "PF" = Pretrained and Frozen. "ViT" = Vision Transformer. "ResNet-50$_{B3, IN}$ = ResNet-50 truncated after the 3rd residual block, with ImageNet transfer learning. "AP" = Attention Pooling only. "GMP" = Global Mean Pooling only. For ease of notion and table space, we remove the subscript which denotes the image resolution operated on by the ViT.

Method	Dim	CRC-100K-R ↑	CRC-100K-N ↑	BCSS ↑	BreasthPathQ ↓
ResNet-50$_{B3, IN}$	1024	0.935	0.983	0.599	0.058
ViT-16$_{PF, BRCA, S1}$	384	0.941	**0.987**	0.593	0.029
ViT-16$_{PF, PANC, S1}$	384	**0.941**	0.983	**0.616**	**0.023**
ViT-16$_{PF, PANC, S4}$	1536	0.927	0.985	0.612	0.052

Table 3. **Patch-Level Classification.** Ablation study assessing K-Nearest Neighbors (K-NN) performance on patch-level datasets with different embedding types. **Abbreviations.** "P" = Pretrained. "F" = Frozen. "PF" = Pretrained and Frozen. "ViT" = Vision Transformer. "ResNet-50$_{B3, IN}$ = ResNet-50 truncated after the 3rd residual block, with ImageNet transfer learning. "BRCA" = Pretrained on BRCA only. "PANC" = Pretrained on Pan-Cancer data. "S1" = Using features from only last stage / hidden layer of ViT. "S4" = Featured concatenated from the last 4 stages. For ease of notion and table space, we remove the subscript which denotes the image resolution operated on by the ViT.

use any Transformer attention in patch- and region-level aggregation, which performed better than HIPT variations with ViT_{4096}-256 finetuned, but still worse than HIPT with ViT_{4096}-256 frozen.

C.2. Assessing Quality of $\mathbf{x}_{256}$ Representations

Though the primary focus of our paper is in hierarchical pretraining using the HIPT architecture, we also make publicly available the pretrained weights of ViT_{256}-16, which can be used as a general feature extractor for 256×256 histology patches. Accordingly, we perform a model audit that assesses the quality of $\mathbf{x}_{256}$ representations for our pan-cancer ViT_{256}-16 model (denoted as ViT_{256}-16$_{PF, PANC, S1}$), and compare with three other embedding types: 1) Pretrained ImageNet (IN) features from a truncated ResNet-50 (after the 3rd residual block, or ResNet-50$_{B3, IN}$), 2) ViT_{256}-16$_{PF, BRCA, S1}$ features trained on only data from TCGA-BRCA (as an organ-specific comparison using the same hyper-parameters and # of iteration), and 3) ViT_{256}-16$_{PF, PANC, S4}$ features trained on pan-cancer data from the TCGA (but with features concatenated across the last four hidden layers, denoted as "S4", versus just the last stage, "S1"). We assess these representations quantitatively using KNN evaluation for slide-level tasks (Table 2) and most patch-level tasks (using global mean pooling (GMP), Table 3), as well as qualitatively using UMAP scatter-plots. Description of patch-level datasets are found below:

- **CRC-100K:** CRC-100K is a dataset of 100,000 histological images of human colorectal cancer and healthy tissue, extracted as 224×224 patches at $20\times$ magnification (available with and without Macenko normalization), and is annotated with the following non-overlapping tissue classes: adipose (Adi), background (Back), debris (Deb), lymphocytes (Lym), mucus (Muc), smooth muscle (Mus), normal colon mucosa (Norm), cancer-associated stroma (Str), colorectal adenocarcinoma epithelium (Tum) [4]. We experiment on CRC-100K with and without Macenko Normalization (denoted with "-N" and "-R" respectively). We report multiclass AUC performance.

- **BCSS**: The Breast Cancer Semantic Segmentation (BCSS) Dataset is a dataset that contains over 20,000 segmentation annotations from the TCGA-BRCA cohort, from which we mined 256×256 patches at $20\times$ magnification for the following overlapping cell types: background tissue, tumor cells, stroma cells, and lym-

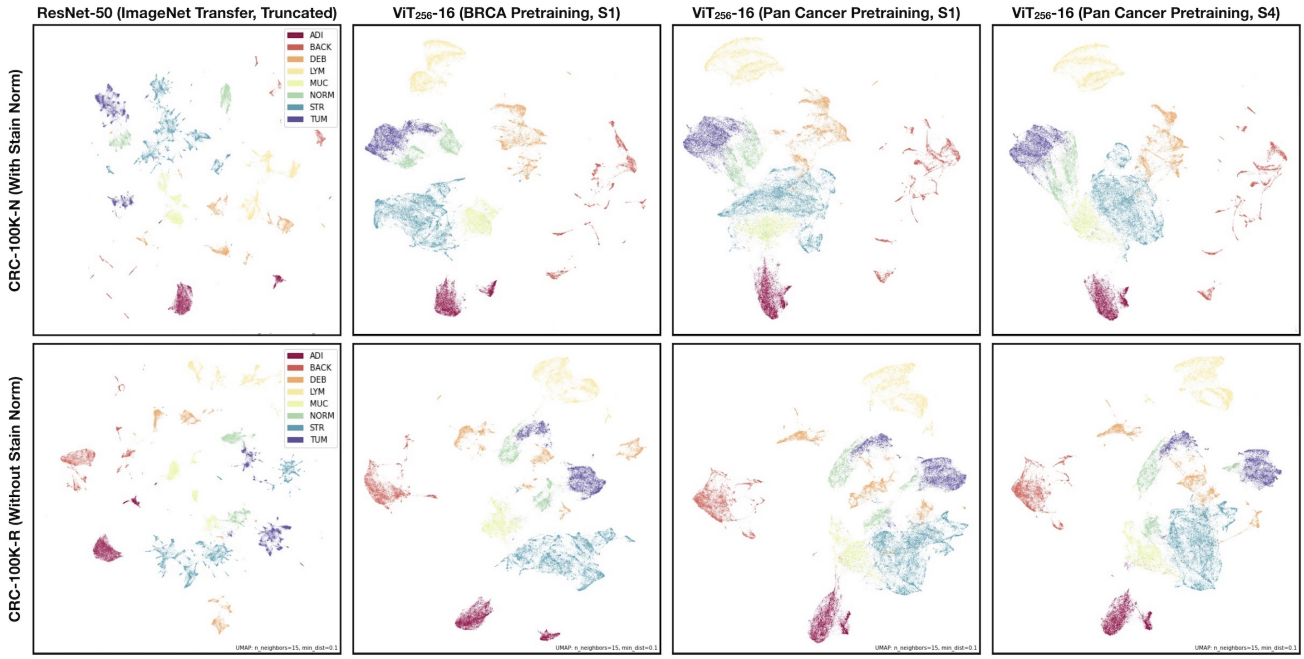

Figure 2. **UMAP Visualization of Pretrained Embeddings**. 2D UMAP scatter-plot visualizing global structure of pretrained embeddings on CRC-100K (with and without Macenko stain normalization). In ResNet-$50_{\text{B3, IN}}$, global structure for many class types are not well-preserved, with worse representation quality in CRC-100K without stain normalization. Across all ViT_{256}-16 models, global structures for morphological subtypes such as normal, tumor, stroma, and mucous are well-preserved in both datasets. We used default UMAP parameters of: $\text{neighbors} = 15, \text{dist} = 0.1$.

phocyte infiltrates [1]. Unlike CRC-100K, BCSS overlaps in label categories as tissue patches can have multiple labels of each cell type. As a result, we used the majority cell phenotype as the patch-level label during supervision. We report multiclass AUC performance.

- **BreastPathQ:** BreastPathQ is a challenge dataset from the TCGA-BRCA cohort that measures tumor cellularity, which measures the fractional occupancy of tumor cell presence in the image patch [6]. We evaluated on the public train/validation split of the challenge, which provides 2579 and 187 patches respectively at $20\times$, and report mean-squared error (MSE) using linear regression.

Comparison with ImageNet Features. In comparing ViT_{256}-$16_{\text{PF, PANC, S1}}$ with ResNet-$50_{\text{B3, IN}}$, on both patch- and slide-level evaluation using KNN, we observe that ViT_{256}-$16_{\text{PF, PANC, S1}}$ features are generally more robust. On patch-level classification, ResNet-$50_{\text{B3, IN}}$ and ViT_{256}-$16_{\text{PF, PANC, S1}}$ do equally well on CRC-100K-N (MNIST equivalent of patch-level tasks in CPATH), with ViT_{256}-$16_{\text{PF, PANC, S1}}$ performing better on BCSS and BreastPathQ (more challenging datasets with noisy and fine-grained labels respectively). ViT_{256}-$16_{\text{PF, PANC, S1}}$ also performed better on CRC-100K-R, in which images are not stain normalized and thus have more variation due to institution-specific staining protocols. One hypothesis for the surprisingly robust ResNet-$50_{\text{B3, IN}}$ features is that the feature maps before the last residual block are low-level feature descriptors, and are able to distinguish between distinct morphologies such as tumor versus stroma, or tumor versus adipose tissue. In visualizing UMAP scatter plots of pre-extracted ResNet-$50_{\text{B3, IN}}$ features, despite the high AUC performance on CRC-100K-R and CRC-100K-N, the representation quality is poor as global structures within the same class types are not preserved (Figure 2). On the other hand, global structures for classes such as stroma, tumor, normal, and mucous tissue are well-preserved for all ViT models.

In slide-level KNN evaluation, interestingly, we see that mean ResNet-$50_{\text{B3, IN}}$ features out-perform mean ViT_{256}-$16_{\text{PF, PANC, S1}}$ features on NSCLC and RCC subtyping, with ViT_{256}-$16_{\text{PF, PANC, S1}}$ performing better on BRCA subtyping. This result can be attributed to NSCLC and RCC subtyping being generally easier tasks in which subtypes can be more readily distinguished, whereas BRCA subtyping being more challenging due to the phenotypic similarity of IDC to ILC that typically requires stroma context.

Organ-Specific versus Pan-Cancer Image Pretraining. In comparing ViT_{256}-$16_{\text{PF, PANC, S1}}$ with

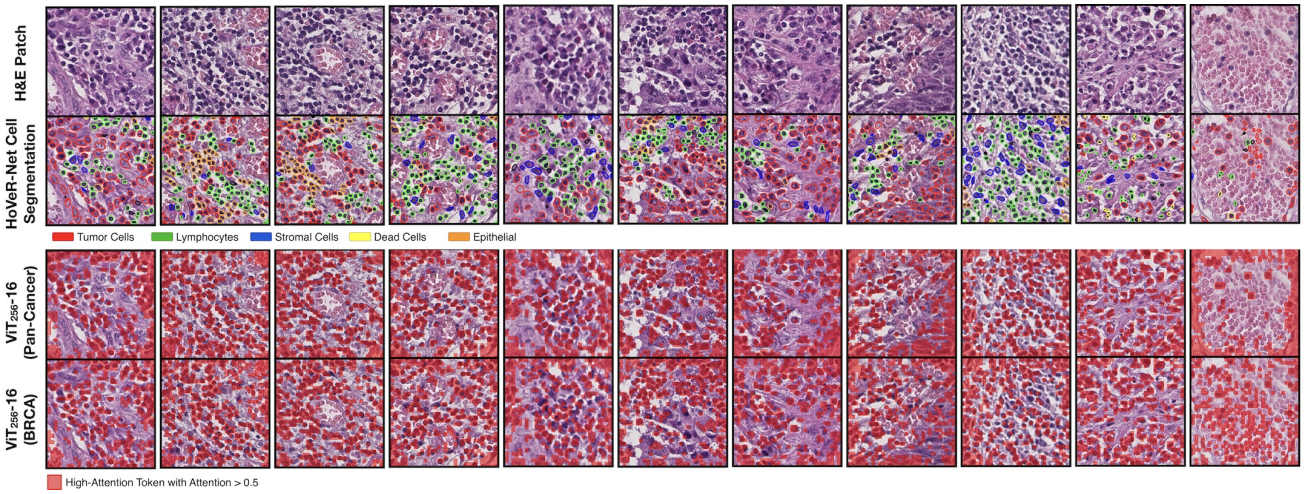

Figure 3. **Comparison of Pan-Cancer versus BRCA pretraining in $\mathbf{ViT_{256}}$-16**. For $\text{ViT}_{256}\text{-16}_{\text{PF, BRCA, S1}}$ and $\text{ViT-16}_{\text{PF, PANC, S1}}$, we visualize the attention weights for $h = 0, 3$ respectively which we observe to be good at localizing cells. For both ViT_{256}-16 models, overlayed in "red" are high-attention visual tokens with attention weights greater than 0.5. In addition, we all show accompanying cell segmentation results from a HoVeR-Net model [3]. Overall, we observe that Pan-Cancer pretraining in ViT_{256}-16 is able to localize tumor and lymphocytes better than BRCA pretraining, with BRCA pretraining attending to blood cells in some instances.

$\text{ViT}_{256}\text{-16}_{\text{PF, BRCA, S1}}$, briefly, we note that both models have similar performance across most patch-level tasks, with $\text{ViT}_{256}\text{-16}_{\text{PF, PANC, S1}}$ performing slightly better on BCSS and BreastPathQ evaluation. In examining global structure of morphological subtypes in CRC-100K, despite not being pretrained on CRC data, $\text{ViT}_{256}\text{-16}_{\text{PF, BRCA, S1}}$ is able to preserve global structure better in stroma and mucous subtypes. In Figure 3, we additionally visualize the cell localization results of each ViT model, with $\text{ViT}_{256}\text{-16}_{\text{PF, PANC, S1}}$ performing overall better in localizing tumor cells and lymphocytes, with $\text{ViT}_{256}\text{-16}_{\text{PF, BRCA, S1}}$ attending more to blood cells.

Feature Concatenation across $\mathbf{ViT_{256}}$-16 Hidden Layers. Following Caron *et al.*, we also concatenated the [CLS] tokens from the last four stages of ViT_{256}-16, resulting in a 1536-dim embedding [2]. We observe no improvement in performing feature concatenation.

Concluding Remarks on $\mathbf{x}_{256}$ Representations. In addition to having strong representation quality and interpretability mechanisms in finding unique morphological features, a key detail about ViT_{256}-16 is that the embedding dimension is relatively small (with a length of 384), which allows Stage 2 Hierarchical Pretraining and finetuning of ViT_{4096}-256 to be tractable on commercial workstations. Though longer embedding dimensions may lead to better performance on some patch-level tasks, our ultimate goal is in building hierarchical models for slide-level representation learning, which relies on: 1) shorter embedding dimensions, and 2) $[16 \times 16]$ cell-level interpretability.

D. Additional Visualizations

Additional visualizations for hierarchical attention maps are shown in Figure 4, 5. Attached in is also a hierarchical attention map for Figure 3 in its native 4096×4096 resolution. Due to space constraints, we created a repository to visualize ViT_{256}-16, ViT_{4096}-256, and hierarchical attention heatmaps at the following link address: https://bit.ly/HIPT-Supplement.

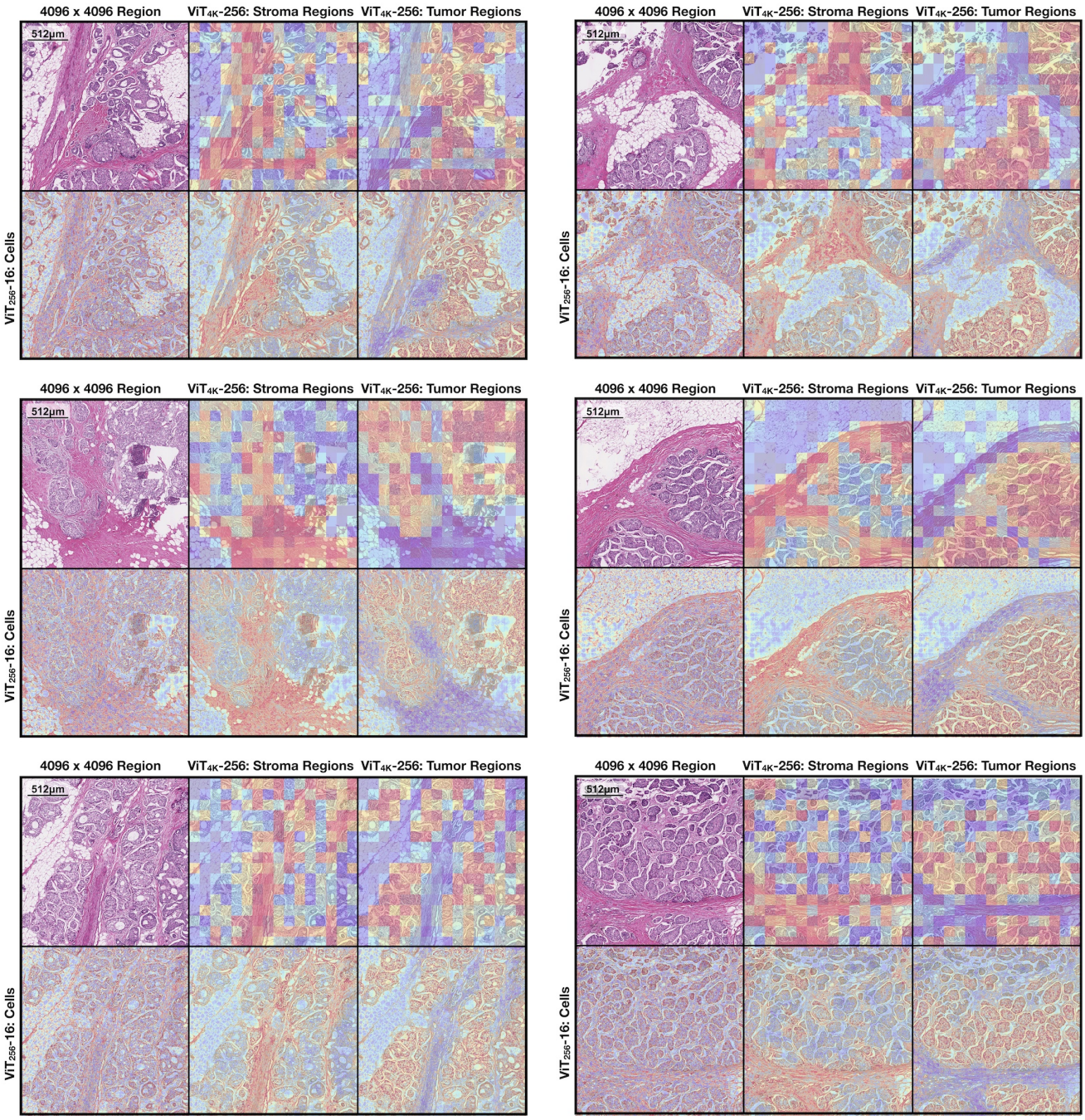

Figure 4. **Hierarchical Attention Maps for Invasive Breast Carcinoma (BRCA)**. Similar to Figure **??**, factorized attention distributions of combined ViT_{256}-16 and ViT_{4096}-256 attention distributions are able to localize: 1) invasive tumor cells in demoplastic stroma, 2) tumor cells arranged in larger tumor nest patterns, which is important in distinguishing Invasive Ductal versus Lobular Carcinoma as well as survival outcomes.

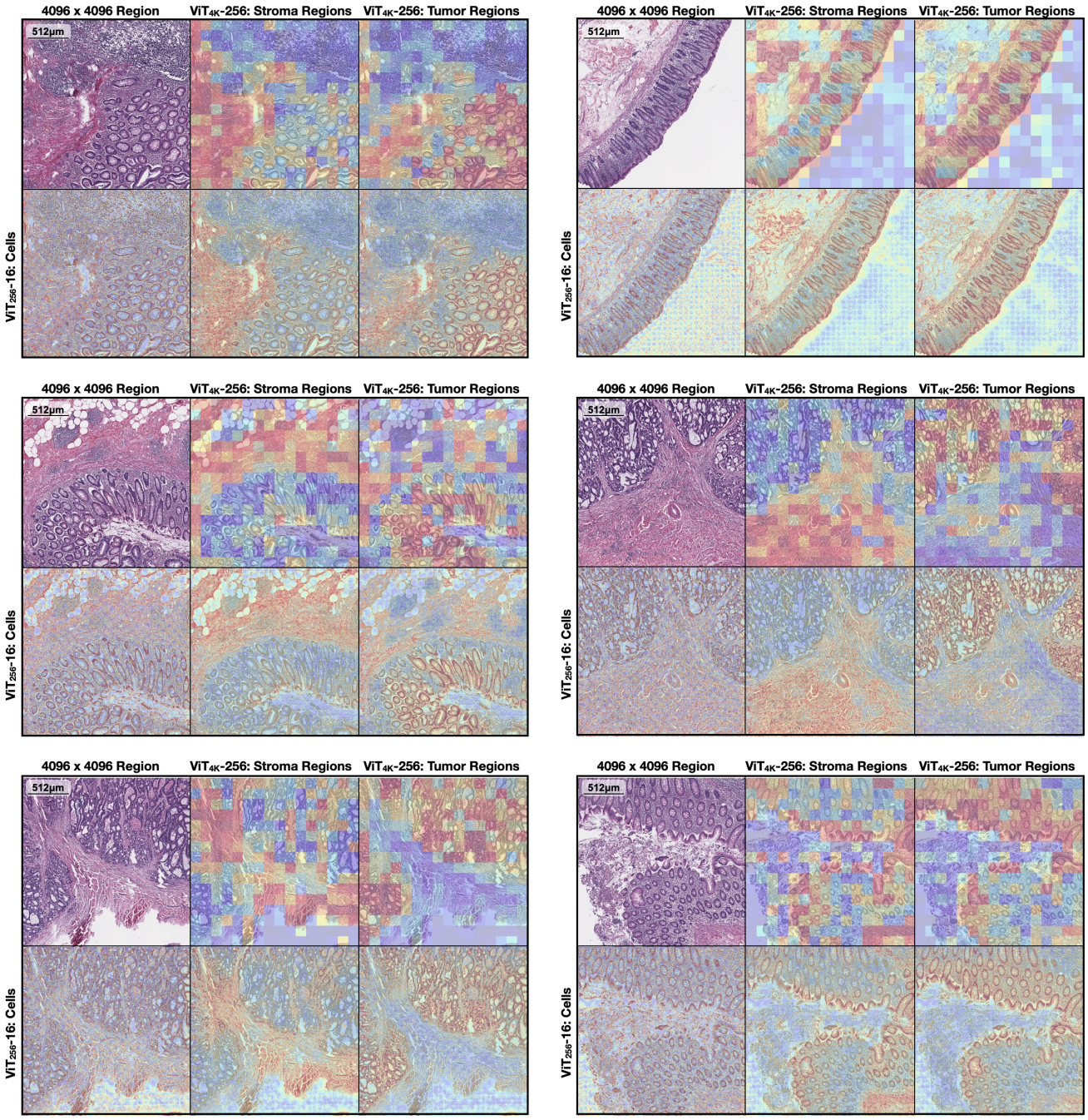

Figure 5. **Hierarchical Attention Maps for Colorectal Cancer (CRC)**. Similar to Figure **??**, factorized attention distributions of combined ViT_{256}-16 and ViT_{4096}-256 attention distributions are able to localize: 1) invasive tumor cells in muscle and stromal regions, 2) tumor cells forming poorly-differentiated glands, which are both important prognostic histopathologic biomarkers in determining severity in cancer staging and survival outcomes.

References

[1] Mohamed Amgad, Habiba Elfandy, Hagar Hussein, Lamees A Atteya, Mai AT Elsebaie, Lamia S Abo Elnasr, Rokia A Sakr, Hazem SE Salem, Ahmed F Ismail, Anas M Saad, et al. Structured crowdsourcing enables convolutional segmentation of histology images. *Bioinformatics*, 35(18):3461–3467, 2019. 4

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*, 2021. 1, 5

[3] Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical Image Analysis*, 58:101563, 2019. 5

[4] Jakob Nikolas Kather, Cleo-Aron Weis, Francesco Bianconi, Susanne M Melchers, Lothar R Schad, Timo Gaiser, Alexander Marx, and Frank Gerrit Zöllner. Multi-class texture analysis in colorectal cancer histology. *Scientific reports*, 6(1):1–11, 2016. 3

[5] Jianfang Liu, Tara Lichtenberg, Katherine A Hoadley, Laila M Poisson, Alexander J Lazar, Andrew D Cherniack, Albert J Kovatich, Christopher C Benz, Douglas A Levine, Adrian V Lee, et al. An integrated tcga pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell*, 173(2):400–416, 2018. 1

[6] Nicholas Petrick, Shazia Akbar, Kenny H Cha, Sharon Nofech-Mozes, Berkman Sahiner, Marios A Gavrielides, Jayashree Kalpathy-Cramer, Karen Drukker, Anne L Martel, et al. Spie-aapm-nci breastpathq challenge: an image analysis challenge for quantitative tumor cellularity assessment in breast cancer histology images following neoadjuvant treatment. *Journal of Medical Imaging*, 8(3):034501, 2021. 4