Learning Where to Look: Pathologist-Inspired Multi-Field-of-View Evidence Retrieval for Cell Type Classification in H&E

Ruizhi Yuan[1], Chongyue Zhao[2], Tianhao Liu[2,3], Qian Wang[4], Zeqiu Yu[5], Lu Tang[1], Heng Huang[6], and Wei Chen[1,2✉]

[1] Department of Biostatistics and Health Data Science, University of Pittsburgh, Pittsburgh, PA, USA
wei.chen@pitt.edu
[2] Department of Pediatrics, University of Pittsburgh, Pittsburgh, PA, USA
[3] Tsinghua Medicine, Tsinghua University, Beijing, China
[4] Department of Pathology, UPMC Children's Hospital of Pittsburgh, Pittsburgh, PA, USA
[5] Department of Electrical and Computer Engineering, University of Virginia, VA, USA
[6] Department of Computer Science, University of Maryland, College Park, MD, USA

Abstract. Accurate classification of cell types in H&E whole-slide images requires integrating evidence across scales, from cell morphology to the local micro-environment and a broader tissue context. However, existing models rely on a fixed field-of-view that frequently misses decisive out-of-view evidence or dilutes local evidence with irrelevant surroundings. We therefore propose a mixture-of-experts (MoE) framework that retrieves multi-field-of-view evidence via query-guided attention pooling, allowing the model to select the right visual cues across views for cell type classification. To ensure that experts effectively collaborate, we deploy an uncertainty-aware router and train the model with a responsibility-driven objective, assigning per-sample credit based on ground truth support to drive complementary specialization across the experts. We validate on public and in-house spatial-transcriptomic-derived datasets of over 400k cell-centered patches, covering lung and skin tissues and diverse cell types, where our method outperforms strong baselines. Finally, the learned pooling weights provide intuitive evidence maps; a small pilot review by a practicing pathologist suggests that these highlighted regions show qualitative agreement with human pathological reasoning.

Keywords: Cell type classification · Digital pathology · Interpretable Deep Learning.

1 Introduction

Cell typing in H&E whole-slide images (WSIs) is now fundamental infrastructure in digital pathology, supporting tasks from tumor microenvironment profiling and spatial statistics to slide-level decision support. With strong nuclei

segmentation baselines (e.g., HoVer-Net and StarDist)[1, 2], large cell-annotated datasets (e.g., PanNuke and Lizard)[3, 4], and modern morphology encoders[5], it is now routine to train models for per-cell classification. Spatial transcriptomics (e.g., Xenium, Visium HD)[6, 7] further scales supervision by providing large quantities of cell labels aligned to H&E, enabling data-efficient training across tissues and cohorts. However, more labels do not automatically resolve the core challenge: choosing the right visual evidence for each cell. Most models still classify cells from a fixed field-of-view (FOV) around each cell. Even vision-transformer(ViT) based pathology foundation models typically encode WSIs as fixed-size tiles (e.g., 224×224 for Virchow, 256×256 for Prov-GigaPath) and leave downstream cell typing to these spatially restricted representations[8–10], which can fail when decisive cues for classification lie outside the chosen FOV.

Fundamentally, if a model is forced to predict from incomplete or irrelevant evidence, even sophisticated architectures will fail. A fixed FOV turns cell typing into a closed-book exam: the model must answer even when the decisive cue lies just outside the window. Pathologists do not work this way—they iteratively consult multiple magnifications, using low power to establish tissue architecture and locate relevant regions, and high power to confirm cellular morphology[11]. Viewport-tracking studies further suggest that such zooming and navigation behaviors are measurable components of real diagnostic workflow[12]. This matters because discriminative evidence for cell typing is inherently multi-FOV and class-dependent: some labels are resolved by cell-intrinsic morphology, while others depend on microenvironmental patterns or broader tissue compartment. Naively enlarging the FOV can introduce irrelevant context and dilute evidence. The WSI literature offers a useful analogy: selective evidence allocation via attention-based aggregation and differentiable zooming can better match diagnostic workflow than uniform processing[13, 14]. Taken together, these observations suggest that the missing piece is not "another scale," but a mechanism that decides when to consult broader context and where to look within each view.

We therefore cast cell typing as multi-FOV evidence retrieval and propose **TRACE**, a mixture-of-experts framework for **T**oken **R**etrieval **A**nd **C**ooperative **E**xperts. In this study, we instantiate TRACE with three experts over nested, cell-centered views: a polygon-masked cell view (in the local crop), a local neighborhood view, and a broader tissue-context view. TRACE leverages pretrained ViT backbones to extract expressive token representations. Unlike recent multi-scale baselines[15], which use view-level representations, our local and context experts perform token-level retrieval: the target cell embedding queries patch tokens via attention to select discriminative regions rather than averaging views. A router uses expert representations and uncertainty cues to produce mixture weights, and fuses expert predictions in probability space. We optimize TRACE with a responsibility-driven objective that aligns routing weights with each expert's posterior support for the ground-truth class, encouraging complementary specialization while discouraging routing collapse. We evaluate TRACE on public and in-house H&E WSIs with cell labels derived from spatial transcriptomics. Unlike conventional H&E annotations that are often limited to fixed-

view patches, spatial transcriptomics provides position-resolved cell-type labels at scale, allowing us to anchor local and context crops at each labeled cell and test instance-specific evidence retrieval across fields of view. On lung and skin cohorts, results show consistent improvements over strong baselines across key cell-typing metrics, with the largest gains on context-dependent and morphologically challenging cases. Finally, a small pilot qualitative review by a practicing pathologist suggests that the retrieved evidence regions are broadly consistent with human pathological reasoning.

2 Method

2.1 Problem setup: multi-FOV evidence for a target cell

Given a detected cell i in a WSI with a polygon P_i and a center $\mathbf{c}_i$, we predict a cell-type label $y_i \in \{1, \ldots, C\}$ for each cell. To capture both cell-intrinsic morphology and contextual cues from the surrounding microenvironment and tissue compartment, we extract two crops centered at $\mathbf{c}_i$: a 224×224 *local* crop and a 1024×1024 *context* crop. We use 224×224 for the local crop, the standard

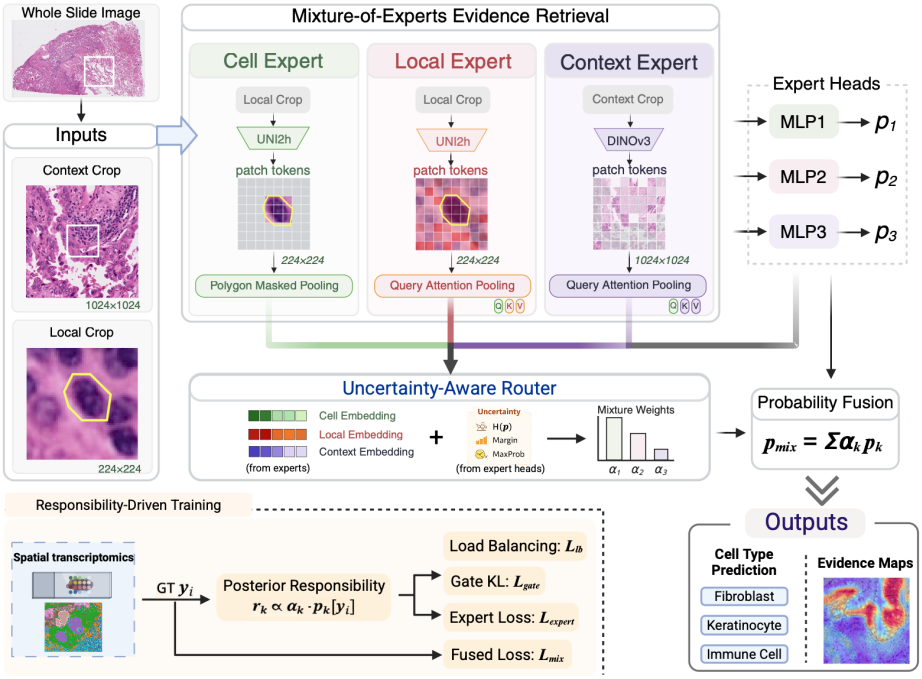

Fig. 1. An overview of TRACE. TRACE retrieves token-level evidence from cell, local, and context views and fuses three expert predictions with an uncertainty-aware router, trained with a responsibility-driven objective.

input size for most pathology encoders, enabling fair comparison and stable feature extraction, and 1024×1024 for the context crop to capture tissue organization while remaining computationally feasible. Our model treats cell typing as instance-specific evidence selection and aggregation across these views.

2.2 Token retrieval and cooperative experts

As illustrated in Fig. 1, TRACE is a MoE framework that, in this study, uses three experts (cell, local, and context) to retrieve token-level evidence from nested views and fuse their predictions via an uncertainty-aware router. We use UNI2h [10], a ViT-based pathology foundation model pretrained on histopathology, to encode the local crop into patch tokens $\{u_{i,j}\}_{j=1}^{N_u}$ with $u_{i,j} \in \mathbb{R}^{D_u}$ for capturing fine-grained cellular morphology; and use DINOv3 [16], a large-scale self-supervised ViT pretrained on diverse natural images, to encode the context crop into $\{v_{i,j}\}_{j=1}^{N_v}$ with $v_{i,j} \in \mathbb{R}^{D_v}$ for capturing broader tissue structure and long-range organization. Here, N_u and N_v denote the numbers of local and context tokens, and D_u and D_v denote their feature dimensions. We retain patch tokens rather than a single pooled embedding, since discriminative evidence in H&E is often sparse and can be diluted by global averaging.

Expert 1 (cell expert): polygon-masked pooling. Cell-intrinsic morphology is often the fundamental evidence, but naive pooling over a local crop can mix in stroma, lumen, or whitespace and corrupt the cell representation. We therefore enforce a simple constraint for the cell expert: compute the embedding only from tokens that fall inside the target cell. Using the provided cell instance polygon P_i (typically a nucleus boundary in H&E), we rasterize it in the local-crop frame and map it onto the UNI token layout, obtaining a binary token mask $M_i = \{m_{i,j}\}_{j=1}^{N_u}$ with $m_{i,j} \in \{0, 1\}$. We then pool only masked tokens:

$$e_{i,1} \;=\; \frac{\sum_{j=1}^{N_u} m_{i,j}\, u_{i,j}}{\sum_{j=1}^{N_u} m_{i,j} + \varepsilon} \in \mathbb{R}^{D_u}. \tag{1}$$

This yields a background-free cell embedding that serves as a stable seed for downstream evidence retrieval.

Expert 2 (local expert): A local crop can contain helpful cues (e.g., immediate neighbors) or substantial irrelevant noise; uniformly pooling the entire view can dilute the few regions that disambiguate the target instance. We therefore aggregate the local view by retrieval: a query derived from the target instance selects informative tokens from the local token set. Let $s_{i,2}$ denote the query source for the local expert; by default $s_{i,2} = e_{i,1}$.

We compute query-based cosine attention over UNI2h tokens. We use cosine attention[17] to make token matching scale-invariant and stable across varying feature norms. We first project the query source and tokens into a shared d-dimensional space:

$$q_i^{(2)} = W_2^{(q)}\,\mathrm{LN}(s_{i,2}), \qquad k_{i,j}^{(2)} = W_2^{(k)}\,\mathrm{LN}(u_{i,j}), \tag{2}$$

then obtain temperature-scaled attention weights and pool the original UNI tokens:

$$a_{i,j}^{(2)} \;=\; \mathrm{softmax}_j\left(\frac{q_i^{(2)} {k_{i,j}^{(2)}}^T}{\tau_2 \,\|k_{i,j}^{(2)}\|\|q_i^{(2)}\|}\right), \qquad e_{i,2} \;=\; \sum_{j=1}^{N_u} a_{i,j}^{(2)}\, u_{i,j}. \tag{3}$$

Using backbone tokens as values (no value projection) makes the attention weights an interpretable evidence trace while preserving the pretrained token space; we learn only query/key matching.

Expert 3 (context expert) We apply the same cosine-attention retrieval to the context tokens $\{v_{i,j}\}_{j=1}^{N_v}$ derived from DINOv3, using $s_{i,3} = e_{i,2}$ by default (or optionally $e_{i,1}$), producing $e_{i,3}$ and attention weights $a_i^{(3)}$. This is implemented by reusing Eqs. (2)–(3) with $(v_{i,j}, W_3^{(q)}, W_3^{(k)}, \tau_3)$.

We project each expert embedding to a shared dimension, $\hat{e}_{i,k} = W_k^{(p)}\,\mathrm{LN}(e_{i,k})$, for routing and classification, and use a small MLP to produce logits $z_{i,k} \in \mathbb{R}^C$.

2.3 Uncertainty-aware routing and probability-space fusion

We fuse expert outputs in probability space to keep training and inference identical and to avoid scale/calibration issues that arise when mixing logits. Let $p_{i,k} = \mathrm{softmax}(z_{i,k})$. The router predicts mixture weights α_i and we form $p_i^{\mathrm{mix}} = \sum_{k=1}^{3} \alpha_{i,k}\, p_{i,k}$. To let the router weight experts by their instance-wise reliability, we augment each expert with simple confidence signals:

$$\phi_{i,k} \;=\; \big[\, H(p_{i,k}),\ \mathrm{margin}(p_{i,k}),\ \max(p_{i,k}) \,\big] \in \mathbb{R}^3, \tag{4}$$

where H is entropy and margin is the gap between the top-1 and top-2 probabilities. Finally, the router conditions on projected expert embeddings and uncertainty cues to produce mixture weights:

$$g_i^{\mathrm{in}} \;=\; [\,\hat{e}_{i,1}, \hat{e}_{i,2}, \hat{e}_{i,3}, \phi_{i,1}, \phi_{i,2}, \phi_{i,3}\,], \qquad \alpha_i \;=\; \mathrm{softmax}(\mathrm{MLP}_g(g_i^{\mathrm{in}})) \in \Delta^2. \tag{5}$$

2.4 Responsibility-driven cooperative training

Which view is decisive varies by instance: some labels are resolved by nucleus morphology, while others require neighborhood or tissue context. Training the router only through the fused prediction provides no explicit credit assignment, which can hinder expert specialization and lead to unstable routing. We therefore use responsibility-driven training: we compute per-sample responsibilities from each expert's support for the ground-truth class, align router weights to these responsibilities, and update experts proportionally.

Fused loss. Let $p_i^{\mathrm{mix}} = \sum_{k=1}^{3} \alpha_{i,k}\, p_{i,k}$. We supervise the mixture by negative log-likelihood (with class weights and label smoothing if used): $L_{\mathrm{mix}} = -\log\big(p_i^{\mathrm{mix}}[y_i] + \varepsilon\big)$.

Posterior responsibilities. Let $p_{i,k}[y_i]$ denote expert k's probability assigned to the ground-truth class. We define posterior responsibilities as

$$r_{i,k} \;=\; \mathrm{stopgrad}\left(\frac{\alpha_{i,k}\, p_{i,k}[y_i]}{\sum_{j=1}^{3} \alpha_{i,j}\, p_{i,j}[y_i] + \varepsilon}\right), \qquad \sum_{k=1}^{3} r_{i,k} = 1. \tag{6}$$

where $\mathrm{stopgrad}(\cdot)$ prevents gradients from flowing through the responsibility targets.

Gate alignment and expert updates. We align the router distribution α_i to r_i via $\mathrm{KL}(r_i \| \alpha_i)$ and train each expert in proportion to its responsibility:

$$L_{\mathrm{gate}} = \frac{1}{B} \sum_{i=1}^{B} \sum_{k=1}^{3} r_{i,k} \Big(\log(r_{i,k} + \varepsilon) - \log(\alpha_{i,k} + \varepsilon) \Big) \tag{7}$$

$$L_{\mathrm{experts}} = \frac{1}{B} \sum_{i=1}^{B} \sum_{k=1}^{3} r_{i,k}\, \mathrm{CE}_{w,\mathrm{smooth}}(z_{i,k}, y_i). \tag{8}$$

Regularization and final objective. To prevent routing collapse (i.e., one expert dominating), we add a standard load-balancing penalty that encourages the batch-average routing to remain close to uniform: $\bar{\alpha}_k = \frac{1}{B} \sum_{i=1}^{B} \alpha_{i,k}$ and $L_{\mathrm{lb}} = \sum_{k=1}^{3} \left(\bar{\alpha}_k - \frac{1}{3}\right)^2$. The overall objective is

$$L \;=\; L_{\mathrm{mix}} \;+\; \lambda_{\mathrm{gate}} L_{\mathrm{gate}} \;+\; \lambda_{\mathrm{exp}} L_{\mathrm{experts}} \;+\; \lambda_{\mathrm{lb}} L_{\mathrm{lb}}. \tag{9}$$

3 Experiments and Results

3.1 Datasets and implementation details

Our evaluation is conducted on 10 H&E WSIs from skin and lung tissue with cell labels derived from aligned spatial transcriptomics: public 10x Xenium[20, 21] and in-house 10x Visium HD. We derive cell-type labels following the official 10x Genomics analysis guides[18, 19]. Across slides there are >400k labeled cells and their paired multi-FOV crops (local/context) centered at cell centroids. We use one in-house WSI and 20% of patch pairs from a held-out contiguous region of the public WSI for validation, and use the remaining data for training.

Experiments are implemented in PyTorch and trained on an NVIDIA H100 GPU. We train TRACE using AdamW with a learning rate of 3e-4, a batch size of 512, mixed-precision training, and early stopping based on validation loss, with a maximum budget of 100 epochs. The UNI2h and DINOv3 backbones are frozen, and only the lightweight query/key projections, expert classifiers, and routing network are optimized jointly. For baselines, we use the official implementations and default hyperparameters, while keeping data splits, preprocessing, and evaluation metrics identical across methods. Code is available at https://github.com/yrzzz/TRACE. Computationally, TRACE avoids full backbone fine-tuning but incurs higher inference cost than fixed-FOV baselines because each target cell requires both local and context-view feature extraction, with DINOv3 encoding of the larger context crop as the main bottleneck.

Table 1. Per-class F1 on lung and skin datasets. We report F1 scores for each cell type across SOTA baselines and TRACE, with the best result per row in **bold**.

Organ	Cell type	Models				
		NuClass[15]	**UNI**[10]	**CellViT**[5]	**HoVer-Net**[1]	**TRACE**
Lung	Airway Epi.	0.90	0.88	0.73	0.68	**0.91**
	Fibroblast	0.07	0.30	0.27	0.22	**0.42**
	Immune	0.70	0.63	0.63	0.60	**0.71**
	Malign. Epi.	0.83	0.83	0.86	0.75	**0.90**
	Normal Epi.	0.02	0.33	0.32	0.31	**0.40**
	Vasculature	0.57	0.45	0.43	0.42	**0.61**
Skin	Endothelial	0.31	0.43	0.33	0.29	**0.44**
	Fibroblast	0.67	0.66	0.49	0.44	**0.69**
	Immune	**0.74**	0.62	0.61	0.48	0.72
	Keratinocyte	0.95	0.92	0.72	0.67	**0.98**
	Malign. Melan.	0.86	0.86	0.80	0.75	**0.90**

3.2 Results

Quantitative Results and Comparative Analysis. Table 1 summarizes per-cell-type F1 on the lung and skin cohorts. We report per-class F1 because cell-type distributions are imbalanced and F1 jointly penalizes false positives and false negatives, making class-specific failures more visible than accuracy alone. Overall, TRACE achieves the best performance among the compared baselines and is notably more robust on morphologically ambiguous, context-dependent cell types. Single-FOV classifiers (UNI[10] with a classifier) attain strong F1 on visually distinctive classes such as Keratinocytes (up to 0.92) and Malignant Epithelium (up to 0.83), but their performance drops substantially on stromal or spindle-shaped categories. For example, on lung Fibroblasts, UNI and CellViT[5] reach 0.30 and 0.27, respectively, and on Vasculature they reach 0.45 and 0.43.

By explicitly retrieving multi-FOV evidence at the token level, our method improves these challenging categories to 0.42 on lung Fibroblasts (+0.12 absolute F1 over UNI) and to 0.61 on Vasculature (+0.18 absolute F1 over CellViT). Compared with the view-level multi-scale baseline NuClass[15], which underperforms sharply on several hard lung classes (e.g., 0.07 on Fibroblasts and 0.02 on Normal Epithelium), our responsibility-driven training maintains discriminative performance (0.42 and 0.40, respectively) without collapsing to majority predictions. The only exception is Skin Immune, where NuClass is slightly higher (0.74 vs. 0.72). Overall, these results support our motivation that adaptive evidence selection across fields of view is particularly important for difficult and context-dependent cell types.

Ablation Study. Table 2 isolates the contributions of token-level retrieval, responsibility-based credit assignment, and uncertainty cues within our framework. Due to class imbalance, accuracy is dominated by frequent classes; we therefore emphasize macro-F1 (mF1). Adding responsibility and uncertainty cues without retrieval slightly improves lung mF1 and maintains skin mF1

Table 2. Comprehensive ablation study of TRACE. We report overall Accuracy and Macro-F1 to highlight the impact of specific modules

Proposed Modules				Lung		Skin	
MoE	**Retrieval**	**Resp.**	**Uncert.**	**Acc**	**mF1**	**Acc**	**mF1**
✓	–	–	–	0.65	0.62	0.76	0.68
✓	–	✓	✓	0.64	0.63	0.73	0.68
✓	✓	–	✓	**0.66**	0.59	0.75	0.67
✓	✓	✓	–	0.64	0.64	0.77	0.70
✓	✓	✓	✓	0.65	**0.66**	**0.77**	**0.71**

(e.g., $0.62 \rightarrow 0.63$ on lung and $0.68 \rightarrow 0.68$ on skin), suggesting that these routing/training cues provide limited benefit without token-level evidence retrieval. In contrast, enabling retrieval without responsibility does not improve mF1 (and can reduce it on lung), indicating that retrieval alone can introduce unstable or noisy expert behavior without explicit credit assignment. The strongest gains occur when retrieval is paired with responsibility, yielding mF1 improvements on both datasets (lung: $0.62 \rightarrow 0.64$, skin: $0.68 \rightarrow 0.70$). Incorporating uncertainty cues in the router provides additional, consistent improvements (lung: $0.64 \rightarrow 0.66$, skin: $0.70 \rightarrow 0.71$), supporting our design choice of combining evidence retrieval with uncertainty-aware fusion and responsibility-driven training.

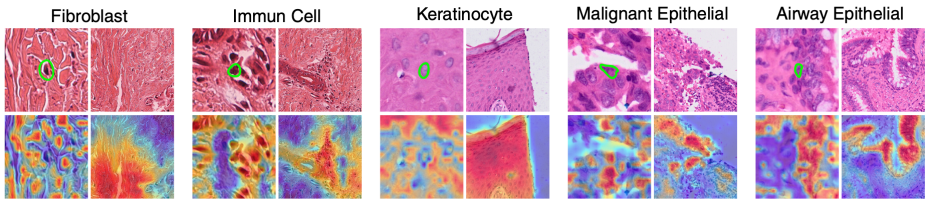

Fig. 2. Evidence maps for five cell types: left/right show local/context views; top shows original crops and bottom evidence maps highlighting regions influential for prediction.

Pilot expert review of evidence maps. TRACE generates evidence maps by reshaping the token-level attention weights from the local and context experts and upsampling them to the crop resolution. We conducted a small pilot review of evidence maps with one practicing pathologist and one medical student on dozens of randomly sampled lung and skin cells. For each cell, we visualized the local and context evidence maps produced by our model (Fig. 2). Both reviewers reported frequent qualitative agreement between highlighted regions and the cues used for cell-type determination, noting that the maps often emphasized cell-intrinsic morphology when sufficient and highlighted broader tissue patterns when additional context was needed. We use this pilot review for evidence auditing and error analysis, rather than as a clinical or causal validation.

4 Discussion

We propose TRACE, a mixture-of-experts multi-FOV framework for H&E cell classification. TRACE casts cell typing as evidence retrieval: a cell expert forms a polygon-masked seed representation, while local and context experts retrieve token-level evidence. On lung and skin cohorts, TRACE improves performance over baselines, especially on context-dependent and ambiguous cell types. A limitation is that evaluation is restricted to lung and skin tissues, so multi-organ generalization remains to be established. In addition, spatial-transcriptomics-derived labels may contain technical noise and should be interpreted as scalable reference labels rather than exhaustive manual ground truth. Future work should evaluate TRACE on larger multi-organ and multi-institutional cohorts to better assess generalizability across diverse histological patterns, incorporate more comprehensive results such as precision-recall analysis and confidence intervals, and develop more efficient context retrieval to reduce compute while preserving evidence quality.

Acknowledgments. This work was supported in part by NIH grants R01AR086434 and R56LM014522, and by the Scleroderma Research Foundation.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Graham, S., Vu, Q.D., Raza, S.E.A., Azam, A., Tsang, Y.W., Kwak, J.T., Rajpoot, N.M.: HoVer-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. Medical Image Analysis **58**, 101563 (2019)
2. Schmidt, U., Weigert, M., Broaddus, C., Myers, G.: Cell detection with star-convex polygons. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11071, pp. 265–273. Springer, Cham (2018)
3. Gamper, J., Alemi Koohbanani, N., Benet, K., Khuram, A., Rajpoot, N.: PanNuke: An Open Pan-Cancer Histology Dataset for Nuclei Instance Segmentation and Classification. In: Reyes-Aldasoro, C., et al. (eds.) Digital Pathology. LNCS, vol. 11435, pp. 11–19. Springer, Cham (2019)
4. Graham, S., Jahanifar, M., Vu, Q.D., Hadjigeorghiou, G., Leech, T., Snead, D., Rajpoot, N.M.: Lizard: A large-scale dataset for colonic nuclear instance segmentation and classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 684–693 (2021)
5. Hörst, F., Rempe, M., Heine, L., Seibold, C., Keyl, J., Baldini, G., Ugurel, S., Siveke, J., Grünwald, B., Egger, J., Kleesiek, J.: CellViT: Vision Transformers for Precise Cell Segmentation and Classification in Histopathology. Medical Image Analysis **94**, 103143 (2024). doi:10.1016/j.media.2024.103143
6. de Oliveira, M.F., Romero, J.P., Chung, M., Williams, S.R., Gottscho, A.D., Gupta, A., et al.; Visium HD Development Team; Taylor, S.E.B.: High-definition spatial transcriptomic profiling of immune cell populations in colorectal cancer. Nat. Genet. 57(6), 1512–1523 (2025). doi:10.1038/s41588-025-02193-3

7. Marco Salas, S., Kuemmerle, L.B., Mattsson-Langseth, C., Tismeyer, S., Avenel, C., Hu, T., et al.: Optimizing Xenium In Situ data utility by quality assessment and best-practice analysis workflows. Nat. Methods 22(4), 813–823 (2025). doi:10.1038/s41592-025-02617-2
8. Vorontsov, E., et al.: A foundation model for clinical-grade computational pathology and rare cancers detection. Nature Medicine (2024)
9. Xu, H., et al.: A whole-slide foundation model for digital pathology from real-world data. Nature **629**, 181–193 (2024)
10. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F.K., et al.: Towards a General-Purpose Foundation Model for Computational Pathology. Nature Medicine (2024)
11. Hipp, J., Monaco, J., Kunju, L.P., Cheng, J., Yagi, Y., Emmert-Buck, M.R., et al.: Integration of architectural and cytologic driven image algorithms for prostate adenocarcinoma identification. Analytical Cellular Pathology **35**(4), 251–265 (2012)
12. Ghezloo, F., Wang, P.-C., Kerr, K.F., Brunyé, T.T., Drew, T., Chang, O.H., Reisch, L.M., Shapiro, L.G., Elmore, J.G.: An analysis of pathologists' viewing processes as they diagnose whole slide digital images. Journal of Pathology Informatics **13**, 100104 (2022)
13. Lu, M.Y., Williamson, D.F.K., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature Biomedical Engineering **5**(6), 555–570 (2021)
14. Thandiackal, K., Chen, B., Pati, P., Jaume, G., Williamson, D.F.K., Gabrani, M., Göksel, O.: Differentiable Zooming for Multiple Instance Learning on Whole-Slide Images. In: European Conference on Computer Vision (ECCV), pp. 77–93 (2022)
15. Xu, Y., Cui, Y., Li, M., Huang, Z.: Adaptive Multi-Scale Integration Unlocks Robust Cell Annotation in Histopathology Images. arXiv preprint arXiv:2511.13586 (2025)
16. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., et al.: DINOv3. arXiv preprint arXiv:2508.10104 (2025)
17. Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., Wei, F., Guo, B.: Swin Transformer V2: Scaling Up Capacity and Resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12009–12019 (2022)
18. 10x Genomics. Annotating Cell Types in Xenium In Situ Data with Label Transfer. Analysis Guide, 2025. https://www.10xgenomics.com/analysis-guides/xenium-cell-type-annotation. Accessed 2026-02-26.
19. 10x Genomics. Introduction to Visium HD Downstream Analysis. Analysis Guide, 2025. https://www.10xgenomics.com/analysis-guides/visium-hd-downstream-analysis. Accessed 2026-02-26.
20. 10x Genomics. Preview Data: FFPE Human Skin Primary Dermal Melanoma with 5K Human Pan Tissue and Pathways Panel. 2024. Licensed under CC BY 4.0. https://www.10xgenomics.com/datasets/xenium-prime-ffpe-human-skin. Accessed: 2026-02-26.
21. 10x Genomics. Post-Xenium Technical Note: Xenium v1 and Xenium Prime 5K for FFPE Human Lung Cancer. 2024. Part of Post-Xenium In Situ Applications Technical Note (CG000709, Rev C). Licensed under CC BY 4.0. https://www.10xgenomics.com/datasets/xenium-human-lung-cancer-post-xenium-technote. Accessed: 2026-02-26.