Cluster-First Labelling: An Automated Pipeline for Segmentation and Morphological Clustering in Histology Whole Slide Images

Muhammad Haseeb Ahmad
Department of Engineering Science
haseeb.ahmad@eng.ox.ac.uk

Sharmila Rajendran
Department of Physiology, Anatomy and Genetics
sharmila.rajendran@dpag.ox.ac.uk

Damion Young
Medical Sciences Division
damion.young@medsci.ox.ac.uk

Jon Mason
Medical Sciences Division
jon.mason@medsci.ox.ac.uk

University of Oxford, United Kingdom

Abstract

Labelling tissue components in histology whole slide images (WSIs) is prohibitively labour-intensive: a single slide may contain tens of thousands of structures—cells, nuclei, and other morphologically distinct objects—each requiring manual boundary delineation and classification. We present a cloud-native, end-to-end pipeline that automates this process through a *cluster-first* paradigm. Our system tiles WSIs, filters out tiles deemed unlikely to contain valuable information, segments tissue components with Cellpose-SAM (including cells, nuclei, and other morphologically similar structures), extracts neural embeddings via a pretrained ResNet-50, reduces dimensionality with UMAP, and groups morphologically similar objects using DBSCAN clustering. Under this paradigm, a human annotator labels representative clusters rather than individual objects, reducing annotation effort by orders of magnitude. We evaluate the pipeline on 3,696 tissue components across 13 diverse tissue types from three species (human, rat, rabbit), measuring how well unsupervised clusters align with independent human labels via per-tile Hungarian-algorithm matching. Our system achieves a weighted cluster–label alignment accuracy of **96.8%**, with 7 of 13 tissue types reaching perfect agreement. The pipeline, a companion labelling web application, and all evaluation code are released as open-source software.

1 Introduction

Whole slide imaging has transformed histopathology by digitising glass slides at high resolution, enabling computational analysis at scale [16, 7]. A critical downstream task is *cell-level annotation*: identifying individual cells, delineating their boundaries, and assigning morphological or functional labels. These annotations provide a valuable educational resource for medical students in our setting.

Terminology. Cell-boundary detection models such as Cellpose segment any morphologically distinct structure that resembles a cell, including individual cells, nuclei, clusters of tightly packed cells, and other tissue components. Filtering these heterogeneous detections at the point of segmentation is impractical without domain-specific heuristics. Throughout this paper we use *cell* as convenient shorthand for any such segmented object; where the distinction matters we write *tissue component* or qualify explicitly.

Manual annotation remains the dominant approach, yet it is prohibitively expensive. A single WSI scanned at $40\times$ magnification can span $100{,}000 \times 100{,}000$ pixels and contain tens of thousands of cells. Tracing boundaries and classifying each cell individually can take days of expert time per slide.

Cluster-first labelling. We propose an alternative paradigm: rather than labelling cells one at a time, our pipeline *first* segments all cells automatically, *then* groups morphologically similar cells into clusters. A human annotator need only review and label each cluster once; the label propagates to every member of that cluster. Because morphologically distinct tissue components—such as nuclei or tightly packed cell groups—are naturally separated into their own clusters, an annotator can label or discard entire categories at the cluster level rather than inspecting individual objects. For example, if a slide contains 15,000 segmented objects grouped into 25 clusters, the annotator reviews 25 representative groups instead of 15,000 individual objects—a reduction of roughly $600\times$.

Contributions.

1. An end-to-end, cloud-native pipeline that takes raw WSI files and produces per-cell cluster assignments, requiring no manual intervention (Section 3).
2. A scalable Azure ML implementation supporting multi-node parallelism with per-slide granularity (Section 3.4).

3. An open-source web application for human validation that computes per-tile Hungarian-aligned accuracy between unsupervised clusters and human labels (Section 3.5).
4. Empirical evaluation on 3,696 cells from 13 tissue types demonstrating 96.8% cluster–label alignment accuracy (Section 4).

2 Related Work

Cell segmentation. Classical approaches relying on thresholding struggle in complex environments, and watershed algorithm struggles with over segmentation. Deep-learning methods such as U-Net [18], StarDist [19], and Hover-Net [9] have significantly improved accuracy. Cellpose [21] introduced a gradient-flow representation that generalises across cell types without retraining. Its latest variant, Cellpose-SAM [15], integrates a Segment Anything backbone to achieve state-of-the-art generalisation, which we adopt in our pipeline.

Computational pathology pipelines. CLAM [12] pioneered attention-based weakly supervised learning on WSIs but operates at the slide level rather than the cell level. QuPath [1] provides interactive cell detection and measurement but requires substantial manual configuration per tissue type. Foundation models for pathology such as UNI [3] offer domain-specific embeddings and are publicly available, albeit behind gated access with restrictive non-commercial licences. Our work is complementary: we combine off-the-shelf, permissively licensed components (Cellpose-SAM, ImageNet ResNet-50) into a fully automated pipeline with an explicit clustering step that enables the cluster-first annotation paradigm. Because the downstream task is unsupervised clustering of relative morphological similarity rather than fine-grained classification, a general-purpose backbone such as ResNet-50 provides sufficient discriminative power while ensuring unrestricted reproducibility.

Unsupervised cell grouping. Extracting meaningful representations of cellular morphology without task-specific labels is a persistent challenge in digital pathology. Unsupervised representation learning has recently shown strong potential for categorizing complex tissue phenotypes [20]. However, many studies focus on tile-level analysis rather than single-cell morphological grouping, or they evaluate clustering on pre-cropped cell datasets rather than bridging the gap to full WSIs [5]. By chaining state-of-the-art segmentation, feature extraction, and clustering into a unified pipeline, we move beyond disjointed tooling. Furthermore, our system pairs unsupervised cell characterisation with a purpose-built human-in-the-loop validation interface that directly consumes pipeline outputs, addressing the practical difficulties of verifying morphological cell clusters.

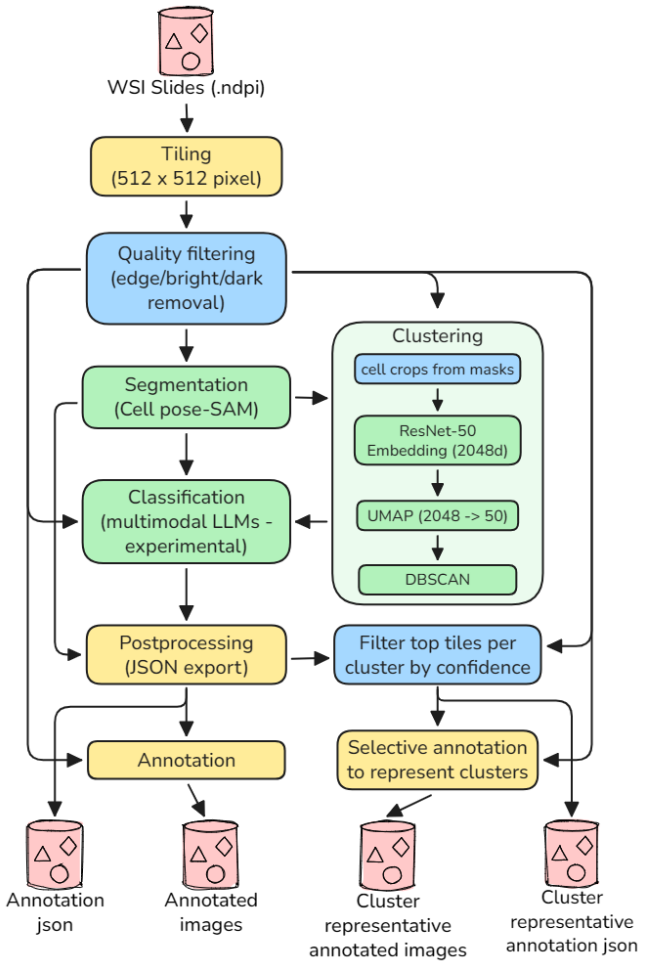

Figure 1: Pipeline architecture. WSI slides (`.ndpi`) are tiled into 512×512 patches, quality-filtered, segmented with Cellpose-SAM, embedded with ResNet-50, reduced via UMAP, and clustered with DBSCAN. Optional downstream stages annotate images, extract representative tiles per cluster, and invoke a multimodal LLM for experimental classification.

3 System Architecture

Figure 1 shows the end-to-end pipeline. Each stage is implemented as an independent, containerised component orchestrated by Azure Machine Learning [14].

3.1 WSI Tiling and Quality Filtering

Raw WSI files are read via OpenSlide [8] and partitioned into non-overlapping 512×512-pixel tiles. The pipeline supports multiple magnification levels via downsampling or upsampling factors relative to the native resolution, and optional uniform sub-sampling to limit the tile count per magnification.

An optional quality-filtering stage removes uninformative tiles using six image-quality metrics: edge density (Canny [2]), bright-pixel ratio, dark-pixel ratio, intensity standard deviation, Laplacian variance [2] (focus quality), and cross-channel colour variance. Tiles failing any threshold are discarded be-

fore segmentation, reducing computation on background and out-of-focus regions.

3.2 Cell Segmentation

We use Cellpose-SAM [15] (`cpsam` model) for boundary detection of cell-like structures. Cellpose-SAM combines the Cellpose gradient-flow formulation with a Segment Anything encoder, achieving robust generalisation across tissue types without fine-tuning. The model detects a variety of objects resembling cells, including individual cells but also nuclei, tightly packed cell groups, and other morphologically distinct tissue components depending on visual context. Discriminating these categories at the segmentation stage is infeasible without tissue-specific filtering rules; instead, the downstream clustering step naturally groups them into separate, visually coherent clusters that can be labelled or discarded at the cluster level.

For each tile, the model produces an instance segmentation mask from which we extract per-object bounding boxes and boundary polygons. Key parameters—flow threshold (0.4), cell-probability threshold (0.0), and optional diameter estimation—are exposed as pipeline arguments to support tissue-specific tuning. GPU acceleration is used by default; tiles can be batched for higher throughput on high-VRAM hardware.

3.3 Neural Embedding and Clustering

Feature extraction. Each segmented cell is cropped from its parent tile using its bounding box and passed through a ResNet-50 [10] backbone pretrained on ImageNet [4]. We chose ResNet-50 for its well-established performance on visual feature extraction and wide availability across deep-learning frameworks. Additionally, because the downstream clustering operates on relative morphological similarity rather than absolute feature quality— the choice of backbone is unlikely to materially affect results. We extract the 2,048-dimensional output of the penultimate average-pooling layer as the cell's feature vector.

Dimensionality reduction. When enabled, UMAP [13] projects the 2,048-dimensional embeddings to 50 dimensions while preserving local and global morphological structure. Default hyper-parameters are $k = 15$ neighbours, $d_{\min} = 0.1$, Euclidean metric.

Clustering. DBSCAN [6] groups cells into clusters without requiring a predefined number of classes. The neighbourhood radius ε is estimated automatically via the knee-point of the sorted k-nearest-neighbour distance curve; the minimum core-point size is set to `min_samples` $= 5$. When a CUDA-capable GPU is available, the RAPIDS cuML [17] implementation is used for acceleration. Cells that do not meet the density criteria are labelled as noise (cluster -1).

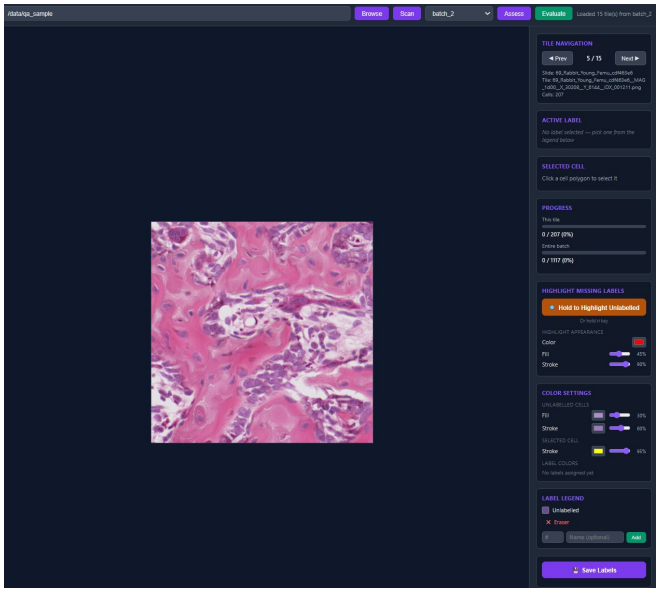

Figure 2: Labelling application interface. The central panel shows a 512×512 histology tile with interactive cell polygons. The sidebar provides tile navigation, label assignment controls, per-tile and batch progress, and visual customisation settings.

3.4 Scalability

The pipeline supports two execution modes selected by a single parameter (`-max_nodes`):

- **Sequential** ($n = 1$): each stage runs as a single Azure ML `command()` job with linear logging, useful for debugging.
- **Parallel** ($n > 1$): stages use Azure ML `parallel_run_function()` with slide-level granularity, distributing slides across up to n GPU nodes.

Both modes execute the same entry-point scripts; only the orchestration wrapper differs.

3.5 Human Evaluation Framework

To measure how well unsupervised clusters align with human judgement, we developed a companion web application (Figure 2). Built with FastAPI, the application:

1. Displays sampled tiles with segmentation-polygon overlays.
2. Lets the annotator define label classes and assign them to cells by clicking on polygons.
3. Exports human labels as timestamped JSON files.
4. Computes accuracy via per-tile Hungarian-algorithm matching [11], then aggregates tile-level results into an overall batch-level accuracy score.

Hungarian-aligned accuracy. Because unsupervised cluster IDs are arbitrary, direct comparison with human labels is meaningless. For each tile, we build a contingency matrix $\mathbf{C} \in \mathbb{Z}^{k \times k}$ where $k = \max(m, h)$, m is the number of model clusters and h the number of human labels, with C_{ij} counting

cells assigned to model cluster i and human label j (entries involving padding rows or columns are zero). The Hungarian algorithm [11] finds the optimal one-to-one mapping π^* from model clusters to human labels that maximises total agreement:

$$\pi^* = \arg\max_{\pi} \sum_{i} C_{i,\,\pi(i)} \tag{1}$$

Accuracy is then the fraction of cells whose mapped model cluster matches their human label. Per-tile alignment is preferred over global alignment because cluster IDs may vary across slides and tiles.

4 Experiments

4.1 Experimental Setup

Data. We processed WSI files (`.ndpi` format) from 13 slides spanning 13 distinct tissue types and three species (Table 2). No slide-specific parameter tuning was performed; the same fixed configuration was used for all tissue types.

Pipeline configuration. Tile quality filtering was enabled using the default thresholds for all six metrics described in Section 3.1. Segmentation used Cellpose-SAM (`cpsam`) with flow threshold 0.4, cell-probability threshold 0.0, and automatic diameter estimation. Embeddings were extracted with a ResNet-50 pretrained on ImageNet and L2-normalised prior to clustering. UMAP reduced dimensionality from 2,048 to 50. DBSCAN used $\varepsilon = 0.5$ and `min_samples` $= 5$; clustering was performed independently per slide rather than globally. GPU acceleration was enabled for both segmentation (Cellpose) and clustering (RAPIDS cuML). The optional LLM classification stage was disabled (`classify_per_cluster` $= 0$). Post-processing aggregated segmentation and clustering outputs into structured per-slide JSON files. Annotated tile images were generated with cluster-ID-coloured polygon overlays. A representative-tile extraction step selected up to 10 cells per cluster with no minimum confidence threshold (`confidence_threshold` $= 0.0$), and a second annotation pass rendered polygon overlays on the filtered representative tiles.

Evaluation. Pipeline outputs were sampled into three evaluation batches using slide-balanced random sampling to ensure every tissue type was represented. Each slide was allocated an equal cell budget ($\lceil 3{,}000/13 \rceil \approx 231$ cells) regardless of its total cell count; tiles within each slide were randomly shuffled and greedily selected until the budget was met. The pooled selection was then shuffled globally and partitioned into three batches of approximately 1,000 cells each. A human annotator independently labelled all cells in each batch tile using the labelling application, without access to cluster assignments. Hungarian-aligned accuracy (Eq. 1) was computed per tile and aggregated per batch.

Table 1: Batch-level evaluation results. Accuracy is the weighted Hungarian-aligned accuracy across all tiles in each batch.

Batch	Cells	Tiles	Human clusters	Model clusters	Acc. (%)
Batch 1	1,229	5	2	1	94.9
Batch 2	1,117	15	3	2	97.7
Batch 3	1,350	9	2	1	97.8
Total	**3,696**	**29**	—	—	**96.8**

Table 2: Per-tissue-type evaluation. Accuracy is the weighted Hungarian-aligned accuracy over all tiles from that slide, pooled across batches.

Tissue type	Species	Cells	Tiles	Acc. (%)
Appendix	Human	662	1	91.4
Ovary	Human	587	2	99.7
Compact bone	Human	106	11	84.0
Skeletal muscle	Human	219	3	84.0
Pancreas	Human	28	1	100.0
Kidney	Rat	328	2	97.9
Prostate	Human	315	1	100.0
Cervix	Human	240	2	100.0
Femur	Rabbit	395	2	99.7
Lung	Human	218	1	100.0
Submandib. gland	Human	178	1	100.0
Seminal vesicle	Human	381	1	100.0
Fallopian tube	Human	39	1	100.0
All tissues	—	**3,696**	**29**	**96.8**

4.2 Results

Batch-level results. Table 1 summarises the three evaluation batches. The pipeline achieves consistently high cluster–label alignment, with a weighted mean accuracy of **96.8%** across 3,696 cells.

Per-tissue analysis. Table 2 breaks down accuracy by tissue type. Seven of 13 tissue types achieve *perfect* cluster–label agreement (100%). Performance is highest on tissues with well-separated, homogeneous cell populations (*e.g.* lung, prostate, cervix) and lowest on compact bone and skeletal muscle, which contain densely packed, morphologically variable structures.

Representative confusion matrices. Figure 3 shows per-tile confusion matrices from the evaluation application. In most tiles the matrix is diagonal or near-diagonal, confirming strong cluster–label alignment. Off-diagonal entries are concentrated in compact bone and skeletal muscle tiles.

Tile-level visual comparison. Figure 4 shows a tile-level post-alignment comparison. Each cell polygon is overlaid with its human label and aligned model label (format: *human / model*). Green polygons indicate matches; red indicates

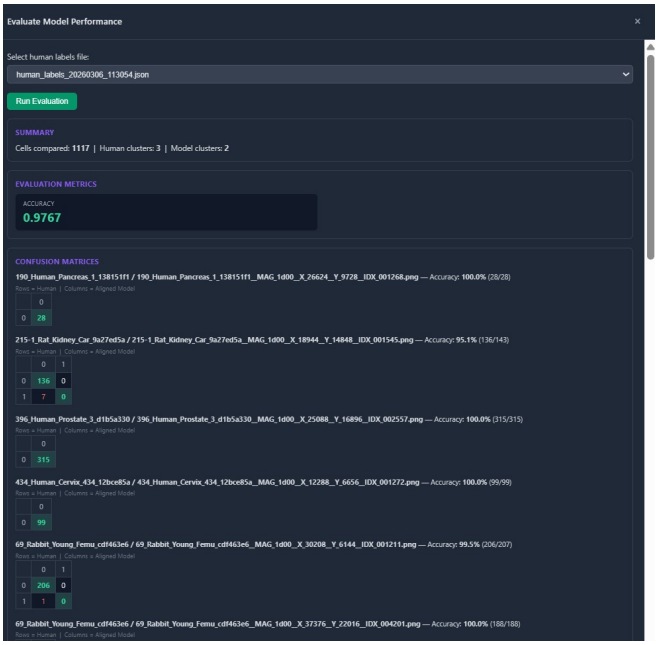

Figure 3: Evaluation interface showing batch-level accuracy (97.67%) and per-tile confusion matrices. Rows represent human labels; columns represent Hungarian-aligned model clusters. Green cells indicate correct assignments.

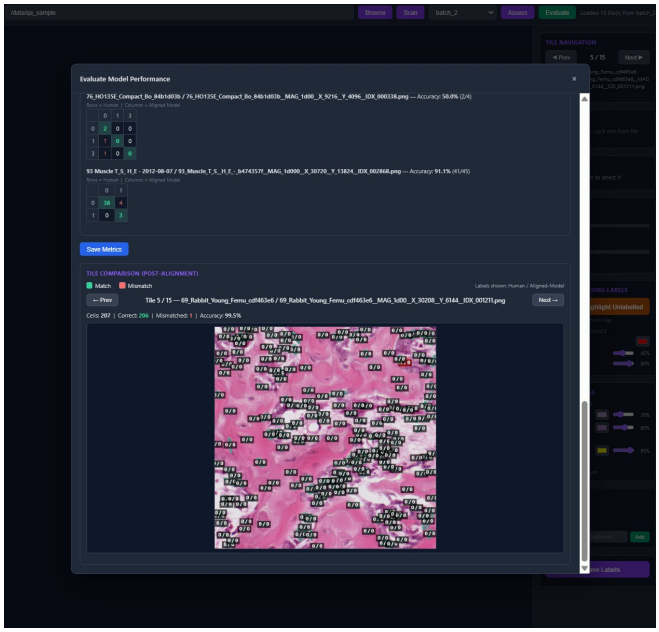

Figure 4: Post-alignment tile comparison. Cell polygons are colour-coded: green = human and aligned-model labels agree; red = disagree. Labels are shown as *human/aligned-model* pairs. Bottom panel: rabbit femur tile with 207 cells, 206 correct (99.5% accuracy).

mismatches. On this tile (rabbit femur, 207 cells), only 1 cell is mismatched, yielding 99.5% accuracy.

4.3 Analysis of Challenging Cases

The two tissue types with accuracy below 90%—compact bone (84.0%) and skeletal muscle (84.0%)—share characteristics that challenge unsupervised clustering:

- **Compact bone** contains very few cells per tile (often < 20), making DBSCAN density estimates unreliable.
- **Skeletal muscle** includes morphologically diverse tissue components—individual muscle fibres, nuclei, and other connective-tissue elements—that a human annotator can distinguish using spatial context. The model, operating on cropped single-object appearance without spatial awareness, may group visually similar but biologically distinct components into the same cluster.

These failure modes suggest that tissue-specific parameter tuning (*e.g.* adjusting ε or `min_samples`) or incorporating spatial context could improve performance on challenging tissues.

5 Discussion

Practical impact. The cluster-first paradigm fundamentally changes the annotation workflow. Instead of $\mathcal{O}(N)$ labelling effort where N is the number of cells, the annotator's effort scales with the number of clusters K, typically $K \ll N$. On our evaluation slides, K ranged from 1–3 human-defined classes; even with DBSCAN producing a modestly higher number of clusters, the annotator's task reduces from reviewing thousands of cells to reviewing tens of clusters.

Generality. Using a single fixed configuration, the pipeline achieves $\geq 97\%$ accuracy on 10 of 13 tissue types and $\geq 99\%$ on 9 of 13. This broad generalisation is enabled by two design choices: (i) Cellpose-SAM's tissue-agnostic segmentation, and (ii) the combination of ResNet-50 feature extraction with UMAP compression, which produces a compact embedding space where morphologically similar objects are nearby and dissimilar objects are well-separated, allowing a single set of DBSCAN parameters to generalise across tissue types. Notably, the pipeline is not restricted to individual cells: it operates on any tissue component detected by Cellpose-SAM, including nuclei, cell clusters, and other structures. The clustering step naturally separates these heterogeneous detections into coherent groups, allowing downstream annotation tools to label or exclude entire categories at the cluster level rather than per object.

Limitations. The evaluation covers 3,696 segmented objects across 29 tiles—sufficient to demonstrate the approach but not exhaustive. Because Cellpose-SAM detects a variety of cell-like structures, the evaluated objects include individual cells, nuclei, and potentially other tissue components; we measure clustering agreement rather than biological classification correctness. Segmentation quality (*i.e.* whether boundaries are pixel-accurate) is outside the scope of this work; we evaluate only whether morphologically similar objects are grouped together. Upstream segmentation errors such as under- or over-

segmentation may propagate to clustering by splitting or merging objects, but this effect is not quantified. The LLM-based classification stage remains experimental and is not evaluated in this work. Performance on tissues such as compact bone or skeletal muscle may benefit from tissue-specific tuning or more spatial awareness.

Open-source release. All pipeline code, the labelling web application, and evaluation scripts are released under the MIT licence at https://github.com/OxfordCompetencyCenters/edu06_histology_labelling (DOI: 10.5281/zenodo.18927168).

6 Conclusion

We presented a fully automated pipeline for cell segmentation and morphological clustering in histology whole slide images, together with a human evaluation framework based on Hungarian-algorithm alignment. Evaluated on 3,696 cells spanning 13 tissue types and three species, the pipeline achieves 96.8% cluster–label alignment accuracy with a single fixed configuration. The *cluster-first* paradigm reduces annotation effort from thousands of individual cells to tens of representative clusters, making large-scale histology annotation practical. Further exploration of integration of spatial context may improve performance on challenging tissues.

Acknowledgements

We thank the University of Oxford, AI Competency Center and the Medical Sciences Division for supporting this research.

References

[1] Peter Bankhead, Maurice B Loughrey, José A Fernández, Yvonne Dombrowski, Darragh G McArt, Philip D Dunne, Stephen McQuaid, Ronan T Gray, Liam J Murray, Helen G Coleman, et al. QuPath: Open source software for digital pathology image analysis. *Scientific Reports*, 7(1):16878, 2017.

[2] Gary Bradski. The OpenCV library. *Dr. Dobb's Journal of Software Tools*, 2000.

[3] Richard J Chen, Tong Ding, Ming Y Lu, Drew F K Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[5] Michael Doron, Théo Moutakanni, Zitong S Chen, Nikita Moshkov, Mathilde Caron, Hugo Touvron, Piotr Bojanowski, Wolfgang M Pernice, and Juan C Caicedo. Unbiased single-cell morphology with self-supervised vision transformers. *bioRxiv*, 2023.

[6] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, pages 226–231, 1996.

[7] Navid Farahani, Anil V Parwani, and Liron Pantanowitz. Whole slide imaging in pathology: advantages, limitations, and emerging perspectives. *Pathology and Laboratory Medicine International*, 7:23–33, 2015.

[8] Adam Goode, Benjamin Gilbert, Jan Harkes, Drazen Jukic, and Mahadev Satyanarayanan. Openslide: A vendor-neutral software foundation for digital pathology. *Journal of Pathology Informatics*, 4(1):27, 2013.

[9] 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.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[11] Harold W Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, 1955.

[12] Ming Y Lu, Drew F K Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6):555–570, 2021.

[13] Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[14] Microsoft. Azure Machine Learning documentation. https://learn.microsoft.com/en-us/azure/machine-learning/, 2026. Accessed: 2026-04-09.

[15] Marius Pachitariu, Michael Rariden, and Carsen Stringer. Cellpose-SAM: superhuman generalization for cellular segmentation. *bioRxiv*, 2025.

[16] Liron Pantanowitz, Paul N Valenstein, Andrew J Evans, Keith J Kaplan, John D Pfeifer, David C Wilbur, Laura C Collins, and Terence J Colgan. Review of the current state of whole slide imaging in pathology. *Journal of Pathology Informatics*, 2(1):36, 2011.

[17] RAPIDS Development Team. RAPIDS: Open GPU data science. https://rapids.ai, 2026. Accessed: 2026-04-09.

[18] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2015*, pages 234–241. Springer, 2015.

[19] Uwe Schmidt, Martin Weigert, Coleman Broaddus, and Gene Myers. Cell detection with star-convex polygons. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 265–273. Springer, 2018.

[20] Chetan L Srinidhi, Ozan Ciga, and Anne L Martel. Deep neural network models for computational histopathology: A survey. *Medical Image Analysis*, 67:101813, 2021.

[21] Carsen Stringer, Tim Wang, Michalis Michaelos, and Marius Pachitariu. Cellpose: a generalist algorithm for cellular segmentation. *Nature Methods*, 18(1):100–106, 2021.