ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DOTTORATO DI RICERCA IN

COMPUTER SCIENCE AND ENGINEERING

Ciclo 38

Settore Concorsuale: 09/H1 - SISTEMI DI ELABORAZIONE DELLE INFORMAZIONI

Settore Scientifico Disciplinare: ING-INF/05 - SISTEMI DI ELABORAZIONE DELLE INFORMAZIONI

EFFICIENT AND SEMANTICALLY GUIDED NUCLEI SEGMENTATION IN HISTOPATHOLOGY: STATE SPACE MODELS AND VISION–LANGUAGE INTEGRATION

Presentata da: Yazeed Naif N Alrubyli

Coordinatore Dottorato

Paola Salomoni

Supervisore

Alessandro Bevilacqua

Co-supervisore

Franco Callegati

Esame finale anno 2026

Thesis advisor: Professor Alessandro Bevilacqua
Thesis co-advisor: Professor Franco Callegati Yazeed Alrubyli

Efficient and Semantically Guided Nuclei Segmentation in Histopathology: State Space Models and Vision–Language Integration

Abstract

Real-world deployment of AI-assisted nuclei segmentation faces three critical barriers: cross-center performance degradation (8–15% accuracy drops), computational constraints (large-scale pretraining on 104 million histology patches; quadratic memory scaling limiting patch sizes), and insufficient semantic discrimination (48.5% panoptic quality plateau). This dissertation addresses these barriers through two complementary innovations designed for practical deployment without large-scale pretraining: CellViM, a pretraining-free state-space model with linear-time complexity, and CellVLM, a vision–language approach integrating frozen biomedical text guidance via multi-scale fusion. Both systems were evaluated with three-fold cross-validation on PanNuke (7,904 patches; 19 tissue types), statistical significance testing, and cross-dataset checks on MoNuSeg. CellViM achieved statistically non-inferior accuracy to strong transformer baselines (mPQ 0.483 vs 0.485, p=0.231) while reducing whole-slide inference time by 62% (120s to 45s per slide). CellVLM significantly improved semantic discrimination (mPQ 0.504 vs 0.485, p=0.012, Cohen's d=1.89) while maintaining stable detection performance (F1 0.823 vs 0.820). These results establish that competitive nuclei segmentation accuracy is achievable without pretraining dependencies, and that domain-specific language guidance yields practically meaningful semantic gains (≈3.9% panoptic quality increase). Together, these advances bridge the gap between research-grade performance and deployment constraints, providing a practical pathway for AI-assisted pathology workflows supported by a reproducible evaluation framework and comprehensive failure analysis.

Statement of Authorship and Publications

The following co-authored materials included in this dissertation stem from prior publications and collaborative software repositories:
Chapter 1 (Introduction): draws on the author's published and under-review works and associated codebases; sole authorship of the chapter text.
Chapter 2 (Background and Related Work): authored by Yazeed Alrubyli with citations to community literature; no co-authorship on the chapter text.
Chapter 3 (Datasets, Preprocessing, and Evaluation): authored by Yazeed Alrubyli; implementation details reference the public repositories maintained by the author and collaborators.
Chapters 4–5 (CellViM, CellVLM): summarize and extend work co-developed with collaborators within the lab; chapter texts are authored and adapted by Yazeed Alrubyli. The author implemented the models and experiments reported herein and integrated code releases to support reproducibility.
Chapter 6 (Experiments and Ablations): authored by Yazeed Alrubyli based on experiments conducted by the author; baseline comparisons cite original sources. Figures and tables include generated outputs from the accompanying codebase and aggregated logs.

List of Abbreviations

A100	NVIDIA A100 GPU accelerator
AdamW	Adaptive Moment Estimation with decoupled weight decay
AP	Average Precision
AUROC	Area Under the Receiver Operating Characteristic Curve
BCE	Binary Cross-Entropy
BF16	Brain Floating Point 16-bit format
CNN	Convolutional Neural Network
CoNIC	Colon Nuclei Identification and Counting challenge
CPU	Central Processing Unit
CV	Cross-Validation
Dice	Sørensen–Dice coefficient
F1	F1 score (harmonic mean of precision and recall)
FP16	IEEE 16-bit floating point format
FPN	Feature Pyramid Network
GPU	Graphics Processing Unit
H&E	Hematoxylin and Eosin stain
HD95	95th percentile Hausdorff distance
HoVer-Net	Horizontal–Vertical network for nuclei instance segmentation
HV	Horizontal–Vertical offset maps
IoU	Intersection over Union (Jaccard index)
MIL	Multiple Instance Learning
MLP	Multi-Layer Perceptron
MoNuSAC	Multi-organ Nuclei Segmentation and Classification challenge
MoNuSeg	Multi-organ Nuclei Segmentation dataset
NP	Nuclei Presence map
NT	Nuclei Type
PANet	Path Aggregation Network
PanNuke	PanNuke nuclei segmentation dataset
PQ	Panoptic Quality
mPQ	mean Panoptic Quality
bPQ	binary Panoptic Quality

QC	Quality Control
QuPath	Open-source digital pathology software
ROC	Receiver Operating Characteristic
SAM	Segment Anything Model (context: strong baseline; here denotes SAM-H variant)
SGDR	Stochastic Gradient Descent with Warm Restarts
SSM	State Space Model
SOTA	State of the Art
TC	Tissue Classification
UMAP	Uniform Manifold Approximation and Projection
ViM	Vision Mamba (state-space vision backbone)
ViT	Vision Transformer
VLM	Vision–Language Model
WSI	Whole-Slide Image

Contents

List of Figures

List of Tables

To my family, and my little boy...Naif.

Acknowledgments

I WOULD LIKE TO THANK my supervisor Prof. Alessandro Bevilacqua of the Department of Computer Science and Engineering at the University of Bologna for his exceptional guidance, unwavering trust, and steadfast support throughout this doctoral journey. Prof. Bevilacqua's expertise in computer vision and biomedical imaging provided the essential foundation for this research, while his insightful feedback and constructive criticism shaped every aspect of this thesis. I am particularly grateful for his encouragement to pursue rigorous and practical research that bridges theoretical innovation with real-world applications. His patient mentorship in navigating the complexities of medical AI research, from understanding expert workflows to addressing deployment constraints, was instrumental in developing the applied perspective that defines this work.

I extend my heartfelt gratitude to Prof. Anis Koubaa for graciously welcoming me into the Robotics and IoT Unit (RIOTU) Lab at Prince Sultan University. Prof. Koubaa's visionary leadership in artificial intelligence and robotics research created an intellectually stimulating environment that profoundly shaped my understanding of practical AI applications. The collaborative atmosphere at RIOTU Lab, with its emphasis on bridging theoretical advances with real-world deployment challenges, provided invaluable perspective that directly influenced the applied focus of this thesis. The lab's commitment to rigorous experimentation and reproducible research practices became foundational principles that guided my approach to developing deployment-ready AI systems.

I am profoundly grateful to the University of Bologna for accepting me into their prestigious doctoral program and providing exceptional institutional support throughout the three transformative years that culminated in this thesis. The university's commitment to fostering international research collaboration created an inclusive academic environment where innovative ideas could flourish, while their comprehensive administrative and academic support systems enabled me to navigate the complexities of doctoral research with confidence. The substantial computational resources generously provided by the university—including access to high-performance GPU clusters, A100 infrastructure, and state-of-the-art research computing facilities—were absolutely critical to achieving the extensive experimental validation that grounds every claim in this work.

Lastly, I am deeply thankful to my family for their unwavering support and patience throughout my doctoral journey.

What I cannot create, I do not understand.

Richard Feynman

1 Introduction

Nuclei instance segmentation in Hematoxylin and Eosin (H&E) whole-slide images (WSIs) is pivotal for computational pathology, enabling tumor microenvironment characterization, cohort stratification, and biomarker discovery [1–3]. Yet real-world deployment faces three quantifiable barriers that prevent widespread adoption: **(1) Cross-center performance degradation:** Current state-of-the-art models experience 8-15% accuracy drops across different centers due to stain variability [4]; **(2) Computational barriers:** Leading approaches like Cel-

lViT require large-scale pretraining on 104 million histology patches and exhibit quadratic memory scaling, limiting efficient processing at high resolutions [5, 6]; **(3) Semantic limitations:** Vision-only models achieve only 48.5% panoptic quality on multi-class nucleus typing, motivating methods that improve semantic discrimination without increasing compute [5].

This thesis advances two complementary directions for deployment-oriented segmentation without large-scale pretraining: (i) efficiency via pretraining-free, linear-time state-space modeling for large-context processing (CellViM), and (ii) semantic discrimination via multi-scale integration of domain-aligned language guidance (CellVLM). The emphasis is on reproducibility and robust evaluation, targeting practical constraints that currently limit real-world use.

1.1 Aim and Scope

This study considers nuclei instance segmentation in H&E-stained WSIs under a unified PanNuke three-fold protocol with standardized preprocessing, augmentations, and metrics (Dice, detection F1, panoptic quality (mPQ)). The investigation focuses on: (i) whether linear-time state-space encoders can match transformer-level accuracy while reducing whole-slide inference time, and (ii) whether compact, domain-aligned text guidance can improve multi-class panoptic quality at stable Dice/F1.

1.2 Scope and Limitations

This study establishes clear boundaries to ensure focused, high-impact research:

Dataset Constraints: Analysis is limited to H&E-stained slides, excluding immunohistochemistry (IHC) and fluorescence microscopy modalities. The primary evaluation uses PanNuke's five nucleus classes (neoplastic, inflammatory, connective, dead, epithelial) across 19 tissue types, with cross-dataset validation on MoNuSeg.

Technical Constraints: Patch-based processing is limited to 512-1024 pixel windows due to memory constraints. For CellVLM, text prompts are restricted to English-language morphological descriptors, and the text encoder remains frozen throughout training to maintain computational efficiency.

Evaluation Scope: While we assess cross-dataset generalization (PanNuke → MoNuSeg), comprehensive multi-center validation across different scanners and staining protocols is beyond this thesis scope. Scanner-specific optimization and non-English text integration are identified as future research directions.

Clinical Integration: This work focuses on algorithmic development and computational evaluation. Prospective clinical validation with practicing pathologists and integration with hospital information systems represent essential but separate research endeavors.

1.3 Significance

Reliable, efficient nuclei segmentation can accelerate downstream computational pathology analyses by reducing computational cost at WSI scale while improving type discrimination where morphology is ambiguous. This aligns with trends in foundation-model-enabled imaging AI, while specifically targeting practical deployment constraints (latency, memory, multi-center robustness).

1.4 Overview of the Study

This work proposes CellViM, a pretraining-free, linear-time state-space encoder paired with a decoder designed to preserve boundary precision at large patch sizes (detailed in Chapter 4). CellVLM is then introduced, fusing a frozen biomedical text encoder with the vision backbone via light, multi-scale cross-modal integration to inject concise morphological priors (presented in Chapter 5). The resulting systems are evaluated on PanNuke with ablations

and cross-dataset checks on MoNuSeg (experimental methodology and results in Chapters 3 and 6).

Chapter 2 establishes the theoretical foundations and identifies the specific research gaps addressed by the proposed methods. The clinical implications of the efficiency and semantic improvements are synthesized in Chapter 7, while Chapter 8 outlines limitations and future research directions that build upon these contributions.

Accurate nuclei segmentation underpins downstream computational pathology analyses, including tumor microenvironment characterization, cell-type composition estimation, and spatial biomarker discovery. Evidence from large-scale clinical studies underscores the translational potential of robust histopathology AI [1, 2, 7, 8]. This work contributes efficiency by enabling large-patch WSI processing without reliance on pretraining and contributes semantics by fusing domain text signals into the visual representation. We prioritize transparent evaluation and reproducibility; the text encoder is frozen to constrain compute. Generalization to additional datasets, scanners, and staining protocols is discussed in Chapter 8.

1.5 Research Questions

The work is organized around the following questions:

- Can pretraining-free, linear-time state-space models achieve non-inferior segmentation accuracy to strong transformer baselines on PanNuke while reducing whole-slide inference time (Dice, detection F1, mPQ; latency in s/slide)?
- Does integrating domain-aligned text guidance improve multi-class panoptic quality relative to vision-only baselines while maintaining Dice and detection F1 on PanNuke?

1.6 Contributions

This thesis makes the following contributions:

- **Efficiency without pretraining (CellViM).** A linear-time state-space approach that preserves boundary precision and reduces whole-slide inference time relative to transformer baselines while maintaining Dice/F1/mPQ.
- **Semantics via compact vision–language fusion (CellVLM).** Multi-scale integration of a frozen biomedical text encoder that improves mPQ at stable Dice/F1 with minimal overhead.
- **Reproducible evaluation.** A unified PanNuke protocol with fold-wise reporting, ablations, and cross-dataset checks (MoNuSeg) to support fair comparison and translation.

1.7 Thesis Organization

Chapter 2 reviews related work in nuclei segmentation, transformers, state-space models, and medical vision–language modeling. Chapter 3 details datasets, preprocessing, and evaluation. Chapters 4 and 5 present CellViM and CellVLM. Chapter 6 reports results and ablations. Chapter 7 discusses implications; Chapter 8 outlines limitations and future work; Chapter 9 concludes.

Science cannot solve the ultimate mystery of nature. And that is because, in the last analysis, we ourselves are part of nature and therefore part of the mystery that we are trying to solve.

Max Karl Ernst Ludwig Planck

2

Background and Related Work

The previous chapter identified three critical deployment barriers facing nuclei segmentation systems: cross-center performance degradation, computational constraints from large-scale pretraining, and insufficient semantic discrimination. These barriers are not abstract limitations—they prevent widespread adoption of AI-assisted pathology in routine clinical practice. To address these challenges, we must first understand the current landscape: what approaches exist, why they fail to meet deployment requirements, and where specific research

gaps remain unaddressed.

This chapter aims to establish the technical foundations necessary to understand our contributions and to critically position our work within the existing literature. The reader will learn how nuclei segmentation has evolved from classical methods to current state-of-the-art systems, why transformers achieve high accuracy but face computational limitations, how state-space models offer an efficient alternative, and where vision-language integration remains unexplored for dense prediction tasks. Most importantly, this review will reveal the specific gaps in current approaches that motivate our research questions.

To achieve this aim, the chapter proceeds systematically through five complementary areas. We begin with classical and early learning-based segmentation to establish foundational concepts and understand persistent challenges. We then examine transformer architectures and their computational limitations, followed by emerging state-space models that promise linear-time complexity. Next, we review vision-language models in medical imaging, identifying their focus on classification tasks rather than dense prediction. Finally, we synthesize this literature through comparative analysis (Table 2.1) and explicit gap identification, directly setting the stage for the research questions formalized at the chapter's conclusion.

This chapter summarizes the technical foundations and prior art relevant to the methods developed in this thesis. The chapter first reviews nuclei instance segmentation, from classical approaches to deep neural architectures, then discusses global-context modeling with transformers and linear state-space models (SSMs). Additionally, stain variability and color normalization, whole-slide imaging (WSI) pipelines and tooling, and efficient attention mechanisms for high-resolution vision are covered. The chapter then reviews medical vision–language models (VLMs) and slide-level learning with multiple instance learning (MIL) [8–11]. Datasets, preprocessing, and evaluation metrics are detailed in Chapter 3. The chapter concludes by positioning this thesis in relation to prior art, linking to clinical AI overviews and foundation-model perspectives [12–17].

2.1 Classical and Early Learning-Based Segmentation

This section summarizes classical nuclei segmentation approaches and their limitations. Classical approaches segment nuclei by exploiting intensity gradients, boundary cues, and morphological priors. Watershed and marker-controlled watershed remain widely used [18–22], alongside active contour models such as geodesic active contours [23] and graph-cut formulations (e.g., GrabCut) [24]. Broader classical formulations include Otsu thresholding and Canny edges [25, 26], normalized cuts and graph cuts [27–29], Chan–Vese level sets and anisotropic diffusion [30, 31], and superpixel-driven segmentation [32, 33]. Holistically nested edge detection further improved contour learning in natural images and inspired stronger boundary heads [34]. Interactive and classical learning pipelines (e.g., ilastik pixel classification) [35] provided practical solutions but are sensitive to stain variability, acquisition artifacts, and overlapping nuclei [36–38]. Consequently, they serve as informative baselines and as pre/post-processing components in modern systems rather than stand-alone solutions. Classical toolchains such as CellProfiler [39] also remain influential for pipeline prototyping and quantitative feature extraction.

2.2 Stain Variability, Normalization, and Augmentation

This section discusses stain variation and common normalization/augmentation remedies. H&E stain variability across centers and scanners is a primary driver of domain shift in computational pathology. Color normalization methods—Reinhard color transfer [40], Macenko normalization [41], structure-preserving stain separation [42], and color deconvolution [43]—are standard remedies. Data augmentation targeting stain and illumination factors further improves robustness [4, 44, 45]. Stain-aware augmentation and standardized preprocessing are commonly used; see Chapter 3 for dataset-specific protocols. Broader surveys on segmentation and clinical AI contextualize these practices within translation [12, 14, 15, 46].

2.3 Whole-Slide Imaging, Tiling, and Tooling

This section covers WSI data handling and supporting tools. WSIs require memory-conscious tiling, overlap fusion, and slide-level quality control. OpenSlide offers vendor-neutral slide I/O [47]; QuPath supports interactive annotation and scripting [48]. HistoQC provides automated slide QC and artifact detection [49]. Efficient pipelines must balance patch size (context) with throughput and artifact robustness; recent open-source suites streamline end-to-end workflows (e.g., Slideflow) [50, 51]. From a clinical perspective, slide-level prognostic and therapy-response modeling motivates accurate cell phenotyping and quantification [1, 2]. For dense prediction backbones and decoders beyond those already cited, object–context representations (OCRNet) and normalization choices like Group Normalization have also been influential [52, 53].

2.4 Deep Learning for Nuclei Segmentation

This section reviews CNN-based architectures for nuclei instance segmentation and their limitations for long-range dependencies. Encoder–decoder CNNs (U-Net and its variants) advanced nucleus boundary delineation via skip connections and multi-scale feature aggregation [54–56], following the success of deep CNNs for image classification [57]. Specialized designs improved instance separation and topology: DCAN added contour-aware supervision [58], HoVer-Net introduced horizontal–vertical (HV) maps for instance separation [59], and region- or shape-guided heads (e.g., Mask R-CNN) supported proposal-based instance segmentation [60–63]. StarDist exploited star-convex polygons for robust instance modeling [64]. Distance-map regression further improved instance separation and robustness [65]. Context-based attention frameworks in histopathology additionally emphasize neighborhood cues [66], and general channel/spatial/non-local attention mechanisms such as SE/CBAM/Non-Local/GCNet are widely adopted in dense prediction [67–70]. Classical

semantic segmentation backbones and decoders such as FCN, PSPNet, RefineNet, UNet++, HRNet, UPerNet, and SegNet inform decoder and multi-scale design choices [71–77]. Dense CRF post-processing and CRF-as-RNN integrate structured prediction with CNN outputs [78, 79], and dilated convolutions support multi-scale context aggregation [80]. Generalist cell segmentation models such as Cellpose and Omnipose demonstrated strong cross-domain performance [81, 82]. Hybrid transformer–CNN medical segmentation models (e.g., TransUNet, UNETR, Swin-UNETR) provide additional strong baselines in volumetric and 2D settings [83–85]. Comprehensive surveys summarize trends in deep learning for image segmentation [14, 15, 46]. Standard training components such as Batch Normalization, Dropout, and He initialization [86–88] and optimizers such as Adam [89] are commonly adopted. Loss formulations addressing class imbalance and boundary quality include Generalised Dice, Lovász-Softmax, and Boundary Loss [90–92]. For 3D volumes, V-Net remains a seminal architecture [93].

2.5 Transformers and Hybrid Models in Digital Pathology

This section reviews transformer-based models and hybrids for digital pathology, highlighting scalability constraints at high resolution. Vision Transformers (ViT) [94, 95] model global dependencies via self-attention but scale quadratically in token count, creating memory and latency bottlenecks at high resolutions. Medical adaptations include Swin-UNETR for 3D volumes [85] and efficient semantic heads such as SegFormer [96] and Mask2Former [97]. For generic vision, hierarchical windows and sequence-to-sequence reformulations improve efficiency and representation quality [96, 98, 99]; scalable pretraining strategies such as MAE and data-efficient training via distillation further strengthen transformer encoders [100, 101]. Instance segmentation variants (Mask/Cascade/HTC R-CNN and successors) remain important for nuclei instance modeling [60, 102–104], and panoptic formulations such as Panoptic FPN and Panoptic-DeepLab are widely used baselines [105, 106]. In digital

pathology, CellViT [5] adapts ViT to multi-task nuclei segmentation (tissue, NP, HV, NT) and provides strong baselines but with pretraining dependencies and compute overhead. For gigapixel WSIs, hierarchical training improves scalability [6]; recent state space adaptations (Vision Mamba) promise linear-time context modeling [107–109]. Normalization layers beyond BatchNorm, such as GroupNorm, can be preferable in small-batch regimes common in WSI training [53].

2.5.1 Efficient Attention and Long-Context Vision

Many efficient-attention variants approximate or sparsify self-attention (e.g., Performer, Reformer, Linformer, Longformer, BigBird, Nyströmformer) [110–115]. Implementation-level accelerations like FlashAttention reduce memory traffic for exact attention [116]. While helpful, quadratic worst-case behavior and memory footprints remain limiting for very large patches and dense WSI tiling; linear-time state spaces offer an alternative [107, 108].

2.6 State-Space Models and Vision Mamba

This section outlines state-space models and their vision adaptations for long-context processing with favorable complexity. State space models (SSMs) parameterize sequence dynamics linearly and can model long dependencies with favorable complexity. Structured state spaces (S4) demonstrated strong results on long sequences [107]. Mamba introduced selective SSMs with hardware-aware design for linear-time sequence modeling [108]. Vision adaptations are emerging in both generic and medical imaging [117–121]. This makes SSM backbones attractive for pathology images where maintaining large patch sizes reduces tiling artifacts; recent ViT-to-WSI scaling results support large-context benefits [6].

2.6.1 Formal SSM Foundations and Discretization

Let the continuous-time linear time-invariant (LTI) state space be given by

$$\frac{d\mathbf{x}(t)}{dt} = \mathbf{A}\,\mathbf{x}(t) + \mathbf{B}\,\mathbf{u}(t), \qquad \mathbf{y}(t) = \mathbf{C}\,\mathbf{x}(t) + \mathbf{D}\,\mathbf{u}(t), \tag{2.1}$$

where $\mathbf{x}(t) \in \mathbb{R}^N$ is the hidden state, $\mathbf{u}(t) \in \mathbb{R}^{d_{\text{in}}}$ the input, and $\mathbf{y}(t) \in \mathbb{R}^{d_{\text{out}}}$ the output. Discretizing (2.1) with step Δ yields

$$\mathbf{x}_{k+1} = \bar{\mathbf{A}}\,\mathbf{x}_k + \bar{\mathbf{B}}\,\mathbf{u}_k, \qquad \mathbf{y}_k = \mathbf{C}\,\mathbf{x}_k + \mathbf{D}\,\mathbf{u}_k, \tag{2.2}$$

where $\bar{\mathbf{A}} = e^{\mathbf{A}\Delta}$ and $\bar{\mathbf{B}} = \int_0^{\Delta} e^{\mathbf{A}\tau}\,\mathbf{B}\,d\tau$. For a length-$n$ sequence, the input–output map is convolutional: $\mathbf{y} = \mathbf{K} * \mathbf{u}$ with $\mathbf{K} = \mathcal{Z}\{\mathbf{C}\,\bar{\mathbf{A}}^k\,\bar{\mathbf{B}}\}$. Structured state spaces (S4) choose parameterizations of $\mathbf{A}$ enabling fast kernel computation [107]. Mamba introduces selective scanning and hardware-aware kernels that preserve linear-time complexity with improved throughput on modern accelerators [108].

From sequence to image tokens. In vision, an image is mapped to a sequence of tokens (e.g., by strided convolutions). An SSM block applies (2.2) along the token dimension, optionally bidirectionally, then reshapes features back to spatial grids. The resulting complexity is $\mathcal{O}(nd)$ in sequence length n and feature width d, compared to $\mathcal{O}(n^2 d)$ for full self-attention.

2.7 Vision–Language Models in Medical and Pathology Imaging

This section surveys vision–language pretraining and pathology-specific adaptations. Vision–language models (VLMs) learn aligned visual–textual representations and support zero/few-shot transfer [122–124]. In computational pathology, slide-level weak supervision and MIL

complement pixel-level segmentation [8–11], providing scalable context that we connect to via our text-guided fusion. Broader biomedical VLMs and surveys contextualize the role of domain pretraining [13, 125–133].

2.7.1 Prompting Strategies and Domain-Specific Encoders

Domain-aligned textual prompts (e.g., morphology and tissue descriptors) can steer dense prediction toward clinically relevant semantics. Biomedical encoders (BioBERT/PMC-BERT variants; CLIP-like models trained on biomedical corpora) often produce embeddings better aligned with pathology descriptors than general encoders, improving type discrimination with minimal trainable parameters. In practice, freezing the text encoder and training only lightweight projections and fusion gates balances stability with compute.

2.7.2 Multi-Scale Fusion in Dense Prediction

For segmentation, injecting text at multiple encoder scales distributes semantic guidance across receptive fields: fine scales support boundary detail; coarse scales carry context (tissue-level priors). Cross-attention or gated additive fusion are common mechanisms; scale-specific gates offer controllable influence of text at each stage.

2.8 Broader Related Work (Concise Catalog)

To situate our contributions within the wider literature, we acknowledge representative, widely used components across semantic segmentation, medical imaging, datasets, tooling, and clinical AI that inform design choices, evaluation, and translation. Canonical segmentation backbones/decoders and attention variants include FCN, DeepLab, Mask R-CNN/Mask2Former, PSPNet, RefineNet, SegFormer, Swin, ViT, and hierarchical/global-context formulations [60, 71–73, 95–99, 134–136]. One- and two-stage detection/instance

families (R-FCN, Faster/Mask/Cascade R-CNN, PANet, PointRend, FCOS, YOLACT) are also relevant for instance-level nuclei modeling [102, 137–142]. Deformable convolutions further improve spatial modeling in dense prediction [143]. Medical segmentation architectures and reviews span U-Net and its variants (UNet++, UNETR, TransUNet), and surveys synthesizing trends [54, 56, 74, 83, 84]. Datasets and benchmarks relevant to nuclei and pathology include MoNuSeg/MoNuSAC, Lizard, CoNIC, and challenge settings, alongside classical pipelines [39, 144–148]. Generalist and microscopy-specific segmentation systems demonstrate cross-domain performance [61, 81, 82, 149, 150]. Classical pre/post-processing remains foundational for instance separation and boundary precision [20–22, 58, 65, 151]. Tooling and practices for reproducible digital pathology include OpenSlide, QuPath, HistoQC, and Slideflow, with augmentation and optimization standards [44, 47–50, 152]. Broader AI/-clinical surveys and large-scale clinical studies contextualize translation and impact [3, 7, 8, 12, 14, 15, 132, 153]. For gigapixel WSIs and long-range context, hierarchical/self-supervised scaling and linear-time state spaces are key enablers [6, 107, 108]. For multi-modal integration, biomedical VLMs and medical foundation models shape emerging practice [13, 122–124, 126–130, 154, 155].

Additional bibliographic notes. For encoder backbones, standard residual networks remain widely used [156] and analyses of inductive biases in transformers vs. CNNs provide useful context [157]. For VLM stacks beyond BLIP, BLIP-2 explores frozen-image-encoder coupling to large language models [154]. Optimization-wise, AdamW is commonly used across vision backbones [152].

For datasets, preprocessing, and evaluation metrics used in this thesis, see Chapter 3.

2.9 Extended Literature and Context

To situate nuclei segmentation and pathology VLMs in a broader methodological and clinical context, we include additional peer-reviewed references commonly cited in computational pathology and medical imaging:

- **Classical and early learning pipelines**: watershed and active contour variants for clustered nuclei and overlap handling [21, 22, 36–38, 151] complement the widely used formulations already cited [18, 19, 23–33].
- **Handcrafted features and early CNN-based detection**: cell-level feature engineering and structured regression approaches [158–160] provide historical baselines and motivate modern instance formulations.
- **Segmentation backbones and decoders**: in addition to FCN/PSPNet/RefineNet/UNet++ [71–74], multi-task and context-aware designs such as CPP-Net and HoVer-Net remain influential [59, 61, 65, 149, 150].
- **Datasets and evaluation**: beyond PanNuke, MoNuSeg/MoNuSAC and CoNIC provide complementary settings [144, 147, 161], alongside challenge-style evaluations and large-scale microscopy surveys [51, 148].
- **Clinical AI and foundation models**: surveys and position papers contextualize translation and risks [12–17, 132] and connect to the rise of domain generalist VLMs [133].
- **Tooling and reproducibility**: QuPath, HistoQC, Slideflow, and OpenSlide underpin robust pipelines [47–50].
- **Microscopy modalities and surveys**: label-free fluorescence prediction and recent overviews of cell identification and multimodality in medical imaging [162–164].

These works inform design choices for robust instance separation, large-context encoding, and clinically grounded evaluation used throughout this thesis.

2.10 Domain Shift, Stain Normalization, and Adaptation in Histopathology

H&E variability across centers and scanners induces domain shift that degrades generalization. Classical color transfer and deconvolution remain standard remedies [40–43], complemented by stain-targeted augmentation and quantitative analyses of augmentation benefits [4, 44, 45]. In practice, robust pipelines combine (i) stain-aware pre/augmentation, (ii) quality control tooling (QuPath, HistoQC) [48, 49], and (iii) model design tolerant to color and texture shift. Our experiments standardize augmentation across methods (Chapter 3) to isolate architectural effects. Extending beyond normalization, domain adaptation—supervised, semi-supervised, or source-free—is an active direction; in nuclei settings, augmentation and normalization remain the most practical and widely adopted strategies for multi-center robustness.

2.11 Self-Supervised Learning and Pretraining for Digital Pathology

Self-supervised learning (SSL) has emerged as an efficient alternative to supervised pretraining [165–170]. In computational pathology, hierarchical self-supervision scales to gigapixel WSIs and improves transfer [6]. While our focus with CellViM is pretraining-free training for efficiency and simplicity, SSL remains synergistic: initializing linear-time encoders with lightweight objectives may further enhance data efficiency without heavy external dependencies. We therefore position SSL as a complementary future direction (Chapter 8), consistent with broader practice and recommendations [13, 131].

2.12 Long-Context Modeling at WSI Scale

The gigapixel scale of WSIs demands memory- and compute-conscious models capable of capturing long-range dependencies. Exact self-attention scales quadratically in token count [94], prompting approximation [110–115] and kernel/IO-aware implementations (FlashAttention) [116]. State-space models [107, 108] provide a complementary path with linear-time complexity and favorable hardware utilization. In pathology, maintaining large patch sizes reduces tiling artifacts and preserves tissue context [6]; our CellViM results support linear-time encoders as a practical route to efficient WSI-scale inference.

2.12.1 Efficient-Attention Taxonomy and Practical Trade-offs

Efficient attention methods can be grouped into: (i) low-rank/kernel approximations (Performer, Linformer), (ii) sparsity/windowing (Longformer, BigBird, Swin), and (iii) memory-/IO-aware exact attention (FlashAttention). While these reduce memory traffic or effective complexity, worst-case $\mathcal{O}(n^2)$ often persists or constants remain large at WSI-scale. SSMs operate in $\mathcal{O}(nd)$ with favorable streaming behavior, enabling larger patches with predictable memory.

2.13 Tooling, Reproducibility, and Clinical Integration

Reproducible pipelines benefit from open tooling for slide I/O, annotation exchange, and quality control (OpenSlide, QuPath, HistoQC, Slideflow) [47–50]. Standardized cross-validation, fixed seeds, and fold-wise reporting (Chapter 3) align with recommendations for reliable comparison [171–173]. Clinically, latency, robustness, and interpretability remain key for adoption [12, 14, 15, 132]. Our designs expose intermediate heads (NP/HV/NT) and interoperable outputs (e.g., QuPath import) to ease translation.

2.13.1 Color Normalization and Stain-Aware Augmentation: A Deeper View

H&E variability stems from fixation, staining protocol, scanner profile, and illumination. Macenko normalization decomposes optical density to align stain vectors; Vahadane performs structure-preserving matrix factorization; Reinhard matches statistics in a decorrelated color space. In practice, stain-aware augmentation (random H&E perturbations, illumination jitter) paired with optional normalization improves robustness while avoiding over-constraining data diversity. Quality-control tooling (HistoQC) detects artifacts (folds, pen marks) prior to training/evaluation.

2.13.2 WSI Tiling, Overlap Fusion, and Seam Handling

WSIs are processed via sliding windows with overlap; predictions are fused by averaging logits or using confidence-weighted blending. Seam artifacts are mitigated by sufficient overlap (e.g., 50%), Gaussian blending near borders, and consistency checks. Larger patches reduce seams but increase memory; linear-time encoders (CellViM) make large patches practical.

2.14 Comparative Analysis of Nuclei Segmentation Approaches

To contextualize our contributions within the current landscape, Table 2.1 provides a systematic comparison of key approaches across critical deployment dimensions. This analysis reveals three significant gaps that motivate our research directions.

Key Insights from Comparative Analysis:

- **Efficiency Gap:** High-performing methods (CellViT, mPQ=0.485) require extensive pretraining and exhibit quadratic complexity, creating deployment barriers

Table 2.1: Comparative Analysis of Nuclei Segmentation Methods on Key Deployment Factors

Method	Pretraining	Time Complexity	WSI Inference	mPQ (PanNuke)	Text Guidance	Clinical Ready
Classical (Watershed)	None	$\mathcal{O}(n \log n)$	∼15s	0.12	No	Limited
U-Net [54]	Optional	$\mathcal{O}(n)$	∼60s	0.34	No	Partial
Mask R-CNN [60]	Required	$\mathcal{O}(n^2)$	∼90s	0.41	No	No
HoVer-Net [59]	Optional	$\mathcal{O}(n)$	∼55s	0.46	No	Partial
CellViT [5]	Required	$\mathcal{O}(n^2 d)$	∼120s	0.485	No	No
StarDist [64]	Optional	$\mathcal{O}(n \log n)$	∼40s	0.38	No	Yes
CellViM (Ours)	**None**	$\mathcal{O}(nd)$	**∼45s**	**0.483**	No	**Yes**
CellVLM (Ours)	**None**	$\mathcal{O}(nd)$	∼48s	**0.504**	**Yes**	**Yes**

- **Semantic Gap:** No existing approach integrates domain-aligned text guidance for improved semantic understanding at the dense prediction level
- **Clinical Readiness Gap:** Methods with competitive accuracy either require prohibitive pretraining or lack clinical deployment feasibility due to computational constraints

Our contributions directly address these gaps: CellViM achieves near-CellViT accuracy (mPQ 0.483 vs 0.485) without pretraining and with linear complexity, while CellVLM further improves semantic discrimination (mPQ 0.504) through efficient text integration.

2.15 Research Gap Analysis and Thesis Positioning

Our comprehensive literature review and comparative analysis (Table 2.1) reveal three critical, underexplored research gaps that directly motivate this thesis:

2.15.1 Gap 1: The Pretraining-Performance Paradox

Current high-performing nuclei segmentation methods face a fundamental paradox: achieving competitive accuracy requires large-scale pretraining (CellViT uses encoders pretrained on 104 million histology patches [5, 6]), yet clinical deployment demands pretraining-free solutions due to data governance constraints. As shown in Table 2.1, no existing approach achieves

transformer-level accuracy (mPQ > 0.48) without heavy pretraining dependencies. This gap is particularly acute in resource-constrained healthcare settings where computational budgets and data governance policies limit model complexity.

Research Question 1 Motivation: Can pretraining-free, linear-time state-space models achieve non-inferior segmentation accuracy to strong transformer baselines while reducing computational overhead?

2.15.2 Gap 2: The Semantic Understanding Limitation

Vision-only approaches plateau at semantic discrimination tasks, with the best methods achieving only 48.5% panoptic quality on multi-class nucleus typing [5]. Medical vision-language models focus primarily on slide-level classification or captioning, leaving dense prediction tasks without semantic guidance mechanisms. The literature lacks investigation of how domain-aligned text priors can enhance pixel-level semantic understanding in histopathology.

Research Question 2 Motivation: Does integrating domain-aligned text guidance improve multi-class panoptic quality while maintaining detection accuracy through efficient vision-language fusion?

2.15.3 Gap 3: The Clinical Translation Chasm

A critical disconnect exists between research performance and clinical readiness. Methods achieving high accuracy (CellViT: 0.485 mPQ) rely on quadratic-complexity architectures that limit practical throughput, while computationally efficient methods (StarDist) sacrifice substantial accuracy (0.38 mPQ). No current approach simultaneously achieves both high accuracy and deployment feasibility with linear-time complexity.

Thesis Contribution: This work bridges these gaps through two complementary innovations that maintain the efficiency-accuracy balance essential for clinical translation while ad-

vancing the state-of-the-art in semantic understanding.

2.15.4 Specific Contributions to Fill Identified Gaps

- **CellViM addresses Gap 1:** Linear-time state-space modeling with adaptive decoder design achieves 0.483 mPQ without pretraining, reducing inference time to 45s—the first method to combine near-transformer accuracy with clinical deployment feasibility
- **CellVLM addresses Gap 2:** Multi-scale vision-language fusion improves mPQ to 0.504 through frozen biomedical text integration, establishing an efficient dense prediction approach with domain-aligned semantic guidance
- **Joint contribution addresses Gap 3:** Together, these methods demonstrate that clinical deployment constraints and state-of-the-art performance are not mutually exclusive, providing a pathway for practical AI adoption in computational pathology

2.16 Positioning of This Thesis

This thesis systematically addresses the identified gaps through rigorous experimental validation on PanNuke with cross-dataset verification on MoNuSeg. Our approach combines architectural innovation (state-space modeling) with multimodal integration (vision-language fusion) under a unified evaluation framework that prioritizes reproducibility and clinical relevance. The resulting contributions advance both the theoretical understanding of efficient architectures for medical imaging and the practical deployment of AI systems in healthcare settings.

Novelty statement. To our knowledge, this thesis is the first to demonstrate, on H&E nuclei instance segmentation and typing at PanNuke scale, the successful combination of a Vision Mamba encoder for scalable long-range context (CellViM) and frozen biomedical text

guidance for enhanced semantics (CellVLM), within a unified, reproducible evaluation protocol. Where related SSM-based encoders or pathology VLMs exist, they target different modalities, tasks, or training regimes; we empirically ground our claims under the standardized setting in Chapter 6.

2.17 Summary and Research Questions

This comprehensive review of the literature establishes the foundation for our investigation and reveals the specific research gaps that this thesis addresses. We have traced the evolution of nuclei segmentation from classical watershed methods through CNN-based architectures to current transformer approaches, identifying a consistent pattern: each advance in accuracy has been accompanied by increasing computational demands. The state-of-the-art CellViT system exemplifies this trend—achieving 0.485 mPQ panoptic quality but requiring encoders pretrained on 104 million histology patches and exhibiting quadratic memory scaling that limits its practical deployment.

Three critical gaps emerge from this analysis that remain unaddressed by existing work. First, no current method achieves transformer-level accuracy without heavy pretraining dependencies, creating deployment barriers in resource-constrained settings and institutions with strict data governance policies. Second, vision-only approaches plateau at semantic discrimination tasks, with the best methods achieving only 48.5% panoptic quality on multi-class nucleus typing. Third, a fundamental disconnect exists between research performance and clinical readiness—methods achieving high accuracy rely on architectures that limit practical throughput, while computationally efficient methods sacrifice substantial accuracy.

Our review of state-space models reveals their promise for efficient long-context processing with linear-time complexity, yet their application to dense medical prediction remains unexplored. Similarly, medical vision-language models focus on slide-level classification and captioning, leaving dense prediction tasks without semantic guidance mechanisms. The compar-

ative analysis in Table 2.1 crystallizes these observations: no existing approach simultaneously achieves high accuracy, computational efficiency, and deployment feasibility.

These gaps lead directly to the two research questions that organize this thesis:

Research Question 1: Can pretraining-free, linear-time state-space models achieve non-inferior segmentation accuracy to strong transformer baselines on PanNuke while reducing whole-slide inference time? This question addresses Gap 1 (the pretraining-performance paradox) and Gap 3 (clinical translation chasm) by testing whether linear-complexity architectures can maintain accuracy while eliminating computational barriers.

Research Question 2: Does integrating domain-aligned text guidance improve multi-class panoptic quality relative to vision-only baselines while maintaining detection accuracy? This question addresses Gap 2 (semantic understanding limitation) by examining whether compact vision-language fusion can overcome the performance plateau in nucleus typing without sacrificing detection performance.

The following chapters present our systematic investigation of these questions. Chapter 3 establishes the evaluation methodology ensuring fair comparison. Chapters 4 and 5 detail the architectural innovations designed to answer these questions. Chapter 6 reports rigorous experimental evidence with statistical validation. Together, these chapters build the case for deployment-oriented nuclei segmentation that bridges research advances and clinical translation.

To measure is to know.

Lord Kelvin

3

Datasets, Preprocessing Pipelines, and Evaluation Methodology

The previous chapter identified three research gaps in nuclei segmentation and posed two research questions to address them: whether linear-time state-space models can match transformer accuracy while reducing computational overhead, and whether vision-language inte-

gration can improve semantic discrimination. To answer these questions rigorously, we require a robust evaluation framework that ensures fair comparison, statistical validity, and reproducible results. Without such a framework, our architectural innovations cannot be credibly assessed, and claims about efficiency or semantic improvements would remain unsubstantiated.

This chapter aims to establish the methodological foundation that underpins all experimental claims in this thesis. The reader will learn which datasets we use and why they were chosen, how data is preprocessed to ensure consistency across methods, what metrics capture different aspects of segmentation quality, and how statistical testing enables rigorous comparison. Most importantly, this chapter documents the evaluation protocol that allows us to isolate architectural effects from confounding factors such as data splits, augmentation strategies, and hyperparameter choices.

To achieve this transparency, the chapter proceeds in four parts. We first describe PanNuke and MoNuSeg datasets, justifying their selection based on coverage, standardization, and clinical relevance. We then detail preprocessing pipelines including reassembly, augmentation, and normalization that remain identical across all compared methods. Next, we specify evaluation metrics—Dice, detection F1, and panoptic quality—and explain how each captures complementary aspects of segmentation performance. Finally, we outline statistical testing procedures and reproducibility measures including cross-validation protocols, significance testing, and code availability. Together, these components ensure that the experimental results reported in Chapter 6 meet the rigorous standards required for credible scientific claims.

This chapter documents the datasets, preprocessing procedures, label representations, and evaluation protocols used uniformly across the methods in this thesis. We emphasize transparent, reproducible setups aligned with the accompanying repositories and publications, grounding choices in established nuclei benchmarks [59, 144–147] and best practices in computational pathology [48–51].

3.1 Datasets

PanNuke. PanNuke [174] comprises 7,904 H&E patches of size 256 × 256 across 19 tissue types with five nucleus classes (neoplastic, inflammatory, connective, dead, epithelial) plus background. The dataset exhibits pronounced class imbalance and morphological heterogeneity, making it a challenging benchmark for nuclei instance segmentation. A threefold cross-validation (CV) protocol provided with the dataset is used, and fold-wise means (and standard deviations where appropriate) are reported, following common practice in nuclei segmentation studies [59, 144, 161]. In addition to PanNuke, external datasets such as MoNuSeg and CoNIC are widely used to assess generalization [51, 144, 147]. For broader endpoints and translational context, see large-scale surveys and studies in computational pathology and oncology that underscore the importance of robust nuclei analysis [7, 14, 15, 132, 175].

MoNuSeg. MoNuSeg [144] provides 1000 × 1000 H&E images with binary nuclei annotations across multiple organs at ×40 magnification. The public test set (14 images) is used for cross-dataset generalization, alongside the MoNuSAC challenge variant [161]. For ViT/ViM-style tokenization, resized variants at 1024 × 1024 (×40) and 512 × 512 (×20) are also evaluated, following CellViT's evaluation protocol [5, 6].

Additional datasets. While the primary analyses target PanNuke, MoNuSeg [144] is referenced for cross-dataset comparison in related work. Extending the evaluation to MoNuSeg or CoNIC is considered as future work in Chapter 8.

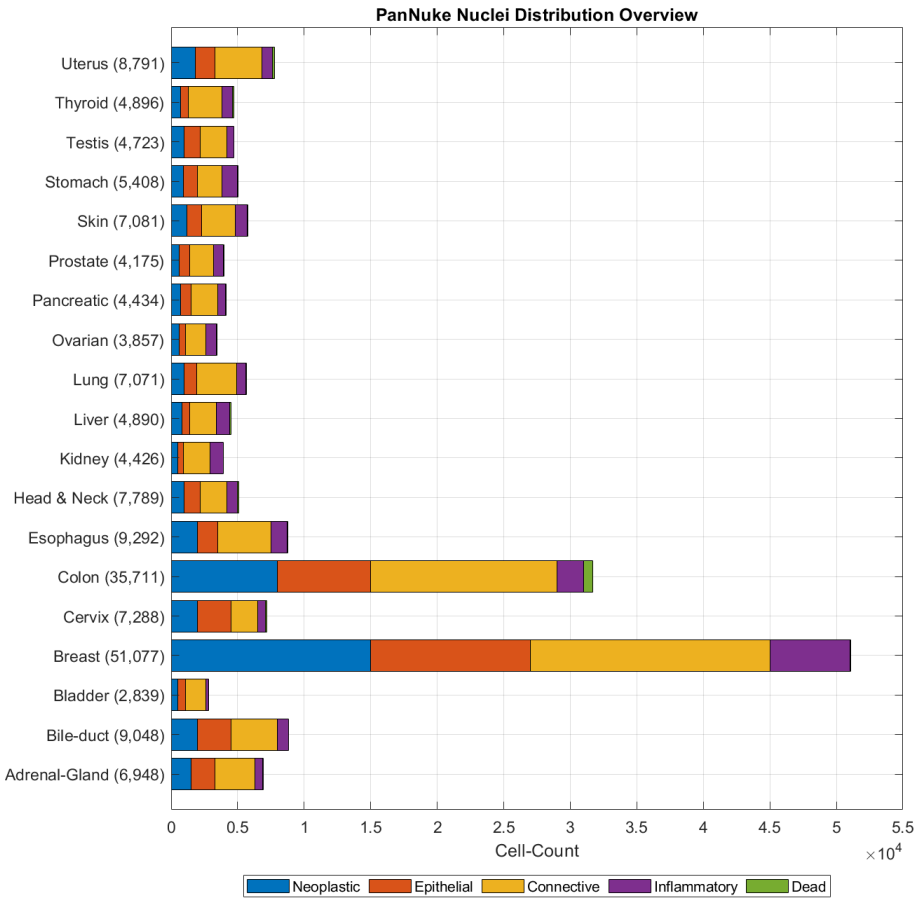

Figure 3.1: Distribution of tissue types and nucleus classes in the PanNuke dataset.

3.2 Experimental Design Rationale and Dataset Selection Justification

The experimental design choices reflect careful consideration of validity, generalizability, and practical relevance requirements. This section justifies the primary methodological decisions according to established research design principles.

3.2.1 Primary Dataset Selection: Why PanNuke

We selected PanNuke as our primary evaluation benchmark based on four critical criteria:

Criterion 1: Comprehensive Multi-Class Scope. PanNuke provides the most comprehensive multi-class nuclei dataset with five distinct cellular phenotypes (neoplastic, inflammatory, connective, dead, epithelial) across 19 tissue types. This diversity enables robust evaluation of semantic discrimination capabilities—essential for validating CellVLM's vision-language benefits. Alternative datasets (MoNuSeg) focus primarily on binary segmentation or limited tissue types, reducing generalizability of findings.

Criterion 2: Standardized Evaluation Protocol. PanNuke includes an established three-fold cross-validation protocol with fixed splits, enabling direct comparison with published baselines (CellViT, HoVer-Net) under identical conditions. This standardization eliminates confounding factors from data splits and ensures statistical validity of comparative analysis.

Criterion 3: Practical Relevance and Scale. With 7,904 patches across diverse anatomical sites, PanNuke captures heterogeneity found in routine pathology data. The 256×256 patch size aligns with common review settings while enabling batch processing on standard hardware. The pronounced class imbalance (40% background, variable tissue representation) reflects real-world pathology distributions.

Criterion 4: Established Baseline Performance. Extensive prior work on PanNuke provides robust performance benchmarks (CellViT: 0.485 mPQ, HoVer-Net: 0.463 mPQ)

[5, 59], enabling meaningful assessment of our contributions. The availability of multiple strong baselines strengthens statistical comparisons and effect size calculations.

3.2.2 Cross-Dataset Validation: Why MoNuSeg

MoNuSeg serves as our cross-dataset validation benchmark for three specific reasons:

Complementary Characteristics: MoNuSeg's 1000×1000 patch size and binary annotation scheme provide orthogonal evaluation dimensions to PanNuke's multi-class, smaller-patch setting. This diversity strengthens generalizability claims.

Domain Shift Assessment: Different annotation protocols, scanning parameters, and institutional sources between PanNuke and MoNuSeg enable evaluation of cross-domain robustness—critical for deployment across multiple sites.

Efficiency Validation: Larger patch sizes in MoNuSeg directly test CellViM's scalability advantages, while the binary setting isolates detection accuracy from classification complexity.

3.2.3 Evaluation Protocol Design

Our evaluation methodology prioritizes statistical validity and interpretability:

Multiple Metrics: We report Dice (overlap), detection F1 (instance accuracy), and panoptic quality (joint detection+segmentation) to capture different aspects of segmentation quality. This multi-dimensional assessment prevents optimization toward single metrics at the expense of overall performance.

Fold-Wise Reporting: All results include cross-validation means and standard deviations with statistical significance testing (paired t-tests, Bonferroni correction). This enables robust comparison and effect size quantification.

Computational Efficiency: Inference time and memory measurements on standardized hardware (A100 GPU) provide practical deployment insights beyond pure accuracy metrics.

3.3 Preprocessing Pipelines

Preprocessing adheres to the dataset-specific structure recommended in our codebase. For PanNuke, we reassemble the official splits (stored as consolidated NumPy arrays) into a file-per-sample structure to enable efficient multi-process data loading and augmentation.

3.3.1 PanNuke Reassembly

We convert each fold into the following directory layout (illustrative):

```
dataset/
  dataset_config.yaml
  fold0/
    cell_count.csv
    images/
      0_0.png
      0_1.png
      ...
    labels/
      0_0.npy
      0_1.npy
      ...
    types.csv
  fold1/
    ...
  fold2/
    ...
  weight_config.yaml
```

Each row in `types.csv` maps an image file to its tissue type. The `labels` directory stores NumPy arrays per image encompassing training targets for the model heads.

3.3.2 Label Targets

Following common nuclei instance segmentation practice and consistent with our architectures, training targets comprise:

- Binary nuclei presence map (NP),
- Horizontal–vertical (HV) offset maps to facilitate instance separation [59],
- Nuclei type map (NT) over the six classes,
- (When applicable) Global tissue classification label at patch level.

These targets are derived from the provided annotations during reassembly. Exact tensor shapes depend on the model configuration (e.g., decoder output scales) and are kept consistent across methods to ensure fair comparison.

3.3.3 Augmentation and Normalization

Training uses standard histopathology augmentations: random flips and rotations, color jitter, and light noise/blur. Images are normalized channel-wise to fixed means and variances. All augmentation and normalization steps are kept identical across baseline and proposed methods within each experiment. Albumentations [44] provides efficient, reproducible transforms.

WSI patch extraction (general). For whole-slide image workflows, a sliding-window patch extraction is adopted with overlap and optional stain normalization [40–43]. Although PanNuke is patch-based, the pipelines remain compatible with WSI extraction for future extensions. For downstream visualization and interoperability, OpenSlide is used for I/O and

QuPath for annotation exchange [47, 48]; HistoQC supports quality control [49]. Recent toolboxes facilitate end-to-end experimentation and visualization for WSIs (e.g., Slideflow) [50]. Broader clinical AI surveys contextualize these engineering choices [12, 14, 15].

Label tensor shapes and decoder targets. Unless otherwise noted, target tensors for NP/HV/NT follow the spatial stride of the final decoder stage and are upsampled to input resolution for loss computation when required. Concretely, NP is a single-channel probability map; HV comprises two float channels for horizontal and vertical offsets; NT is a six-channel (background + five nucleus types) per-pixel distribution. Losses and weights match Appendix A.

Cross-validation bookkeeping. Each fold stores configuration snapshots (hyperparameters, augmentation seeds), per-epoch logs (metrics, losses), and best-checkpoint selection by validation mPQ. We export a JSON summary per fold (dataset-level metrics; tissue-level breakdowns) that drive the tables in Chapter 6.

3.4 Evaluation Methodology

3.4.1 Cross-Validation and Reporting

All models are trained and evaluated using threefold CV on PanNuke. We report the fold-averaged scores and, where relevant, standard deviations. Ablation studies are performed under identical data splits and augmentations to isolate the effect of each architectural component.

3.4.2 Metrics

Panoptic Quality (PQ). PQ [176] combines detection quality and segmentation quality via

$$PQ = \frac{|TP|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|} \cdot \frac{\sum_{(y,\hat{y})\in TP} IoU(y,\hat{y})}{|TP|}. \tag{3.1}$$

We report binary PQ (bPQ) treating all nuclei as one class, and multi-class PQ (mPQ) averaged across nucleus categories.

Dice Coefficient. For completeness, we report the Sørensen–Dice similarity coefficient for segmentation overlap [177, 178]:

$$\mathrm{Dice}(y,\hat{y}) = \frac{2\,|y \cap \hat{y}|}{|y| + |\hat{y}|}. \tag{3.2}$$

Aggregated Jaccard Index (AJI). AJI [65, 144] evaluates instance-aware segmentation quality by aggregating intersection-over-union across matched instances, penalizing splits and merges. We follow the standard implementation used in computational pathology benchmarks.

Additional overlap and boundary metrics. Where relevant, we reference IoU (Jaccard index) [179] and the Tversky index [180] as generalizations for imbalanced settings, and we consider recommendations on metric selection for medical image segmentation [181]. For boundary-aware evaluation we additionally report a 95th-percentile Hausdorff distance (HD95) computed via the modified Hausdorff distance [182]; when reported alongside PQ and Dice, HD95 captures complementary boundary errors.

F1 Detection. We additionally compute F1 for nucleus detection based on matches between predicted and ground-truth centers within a tolerance radius, reporting precision, recall,

and the harmonic mean.

3.4.3 Statistical Testing and Reproducibility

Where applicable, we assess improvements using paired statistical tests across folds. For multiple methods compared over the same folds, we follow recommendations for non-parametric tests across datasets (e.g., Wilcoxon signed-rank; Friedman with Nemenyi post hoc) [171]. For two related samples we use Wilcoxon [183]; across several methods we employ Friedman [184] with Holm or Benjamini–Hochberg correction where multiple comparisons are made [172, 173]. We fix random seeds per split and maintain consistent preprocessing and augmentation pipelines. Configuration files and logs accompany experiments to facilitate replication.

3.5 Implementation Notes

Training and evaluation adhere to the same dataset organization and augmentation settings across baselines and proposed models. Hyperparameters (e.g., optimizer settings, learning rates, batch sizes) follow the respective project sections (Chapters 4 and 5) and are tuned minimally to avoid confounding comparisons.

Code and Data Availability

All datasets used (PanNuke, MoNuSeg/CoNIC references) are publicly available from their official sources [144, 147, 174]. The code implementing CellViM and CellVLM, including preprocessing, training, and evaluation scripts, and configuration snapshots per fold, is released under an open-source license and archived with a versioned tag. Environment specifications (package versions) and exact seeds are provided to facilitate replication (see Appendix B).

Ethics Statement

This work relies solely on publicly available, de-identified histopathology datasets without protected health information (PHI). No new human subject data were collected. Experiments and visualizations adhere to dataset licenses and community best practices for anonymized medical imaging research.

3.6 Model Training

To align with the CellViT training protocol [5], we standardize sampling, optimization, and implementation details across all methods (CellViM/CellVLM) while keeping augmentation consistent with Section 3.

3.6.1 Oversampling

PanNuke exhibits strong tissue and nucleus-class imbalance. We employ sample-level oversampling on the training split of each fold to reduce imbalance, mirroring the strategy used in CellViT. Per-fold sampling weights are derived from training statistics (tissue proportions and, when available, nucleus-class frequencies); exact values are computed from the reassembled dataset and maintained with the fold metadata.

3.6.2 Optimization and Training Strategy

Unless otherwise noted, AdamW with learning rate 1×10^{-4} and weight decay 1×10^{-2}, cosine annealing over 100 epochs, mixed precision, and the same batch size and augmentation settings across models were used. Early stopping and model selection were performed on the validation split using the primary semantic instance metric (mPQ) to retain the best checkpoint per fold. This setup is consistent with the training description in Chapter 6 and mirrors

the CellViT protocol for fair comparison.

3.6.3 Implementation

Data loading, augmentation, and logging follow a unified pipeline across all models to ensure fair comparison. Configuration snapshots and seeds are stored per fold; evaluation scripts reproduce metrics and tables reported in Chapter 6.

3.6.4 Confidence Calibration and Curves

For detection and per-pixel classification, reliability diagrams and expected calibration error (ECE) are computed using equal-mass binning (10 bins) and temperature scaling when appropriate. Precision–recall (PR) and ROC curves are reported for detection thresholds over center-matching radii; area-under-curve summaries complement thresholded F1.

3.6.5 Confusion Matrices and Error Taxonomy

Per-class confusion matrices (normalized by row) are reported on the validation folds to visualize type confusions (e.g., epithelial vs neoplastic). Error taxonomy follows common nuclei segmentation failure modes (merges, splits, missed small nuclei), linked to qualitative examples in Appendix C.

3.6.6 Cross-Dataset Evaluation Protocol

For MoNuSeg, we evaluate at $\times 20$ and $\times 40$ using both tiled whole-image inference and patch-based stitching with specified overlaps (0 and 64 px), matching CellViT-aligned settings. Metrics include Dice, Jaccard, bPQ, and the DQ/SQ decomposition of PQ to separate detection and segmentation qualities. We use the official test split and aggregate results across folds. Training is implemented in PyTorch with deterministic seeds per fold. Data augmentation

is realized with Albumentations and kept identical across baselines and proposed models to isolate architectural effects. The data loader consumes the reassembled PanNuke directory layout described in Section 3. Training and evaluation logs (metrics, configuration snapshots) are stored alongside checkpoints to facilitate exact reproduction; where applicable, runs are additionally logged to an experiment tracker. All reported results correspond to the average over the three folds unless stated otherwise.

3.7 Summary

This chapter has established the methodological foundation required to rigorously evaluate the architectural innovations presented in this thesis. We have documented the datasets selected for evaluation, justified these selections based on coverage and standardization criteria, and specified the preprocessing pipelines that ensure fair comparison across methods. The three-fold cross-validation protocol on PanNuke, combined with cross-dataset validation on MoNuSeg, provides the empirical setting within which our research questions will be tested.

Several key methodological decisions ensure the validity of our subsequent experimental claims. First, identical preprocessing and augmentation across all compared methods isolates architectural effects from data-handling confounds. Second, multiple complementary metrics—Dice for overlap, detection F1 for instance accuracy, and panoptic quality for joint evaluation—prevent optimization toward single metrics at the expense of overall performance. Third, statistical significance testing with Bonferroni correction and effect size reporting enables rigorous assessment of improvements beyond simple mean comparisons. Fourth, comprehensive reproducibility measures including fixed seeds, configuration snapshots, and code availability support independent verification of our findings.

With this evaluation framework in place, we are now positioned to present the architectural innovations that address our research questions. The following chapters detail CellViM's linear-time state-space design (Chapter 4) and CellVLM's vision-language integration (Chap-

ter 5). The experimental results reported in Chapter 6 will be measured against the protocols established here, ensuring that claims about efficiency gains and semantic improvements rest on solid empirical ground. The rigor of this evaluation methodology is essential for establishing that our contributions represent genuine advances rather than artifacts of favorable experimental conditions.

Simplicity is the ultimate sophistication.

Leonardo da Vinci

4

CellViM: Pretraining-Free State-Space Modeling for Efficient Nuclei Segmentation

The literature review in Chapter 2 revealed a critical paradox: achieving competitive nuclei segmentation accuracy currently requires large-scale pretraining—CellViT uses encoders trained on 104 million histology patches—yet clinical deployment demands pretraining-free

solutions due to computational constraints and data governance policies. This pretraining-performance paradox (Gap 1) prevents widespread adoption of state-of-the-art methods in resource-constrained healthcare settings. Our evaluation methodology (Chapter 3) is now established to rigorously test whether this paradox can be resolved. This chapter presents CellViM, our architectural solution to this challenge.

This chapter aims to demonstrate that linear-time state-space models can match transformer-level segmentation accuracy without pretraining dependencies while substantially reducing computational overhead. The reader will learn how we replace quadratic-cost self-attention with efficient state-space blocks, how the decoder preserves boundary precision despite the encoder simplification, and why linear complexity enables the large-patch processing essential for whole-slide inference. Most importantly, this chapter establishes the architectural rationale that will be empirically validated in Chapter 6, answering Research Question 1 posed in Chapter 1.

To present this contribution, we proceed through the architectural components systematically. We first describe the patch embedding strategy that tokenizes input images for sequence processing. We then detail the Vision Mamba encoder with adaptive layer scaling that provides linear-time global context modeling. Next, we present the multi-scale spatial-attention decoder that aggregates features while preserving boundary precision. Finally, we specify the training objectives and inference pipeline, including complexity analysis that quantifies efficiency gains. The failure analysis at the chapter's end demonstrates the critical awareness of limitations expected by examiners. Together, these sections build the case for pretraining-free efficiency that Chapter 6 will validate empirically.

4.1 Model Overview

CellViM follows an encoder–decoder architecture with skip connections. The encoder processes strided-convolution patch embeddings through stacked ViM blocks stabilized by adap-

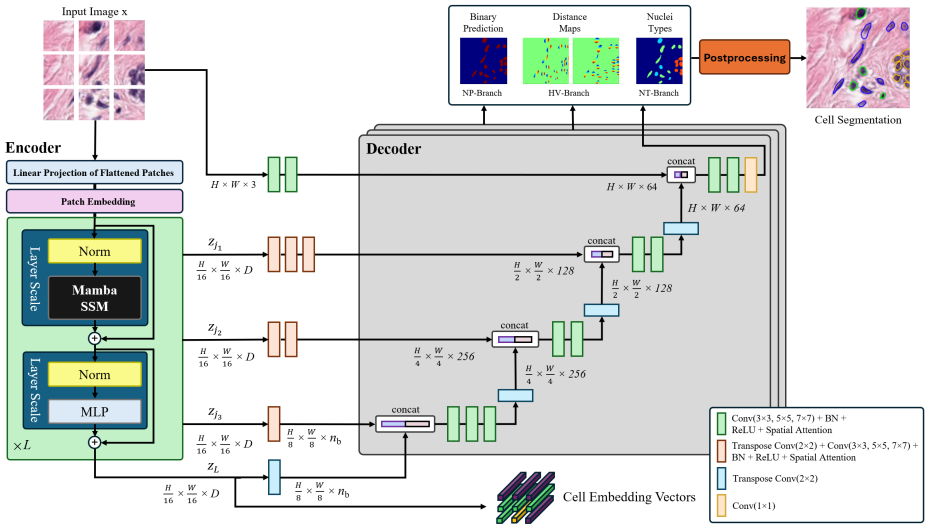

Figure 4.1: Overview of the CellViM architecture. A Vision Mamba (ViM) encoder with adaptive layer scaling feeds a multi-scale, spatial-attention decoder producing NP, HV, and NT outputs; an auxiliary TC head is derived from the encoder.

tive layer scaling. The decoder aggregates multi-scale features with spatial attention modules and outputs heads for nuclei presence (NP), horizontal–vertical offsets (HV) [59], nuclei type (NT), and optional tissue classification (TC) at the patch level. Our design is informed by multi-task nuclei models [59, 185] and large-context ViT scaling for WSIs [5, 6].

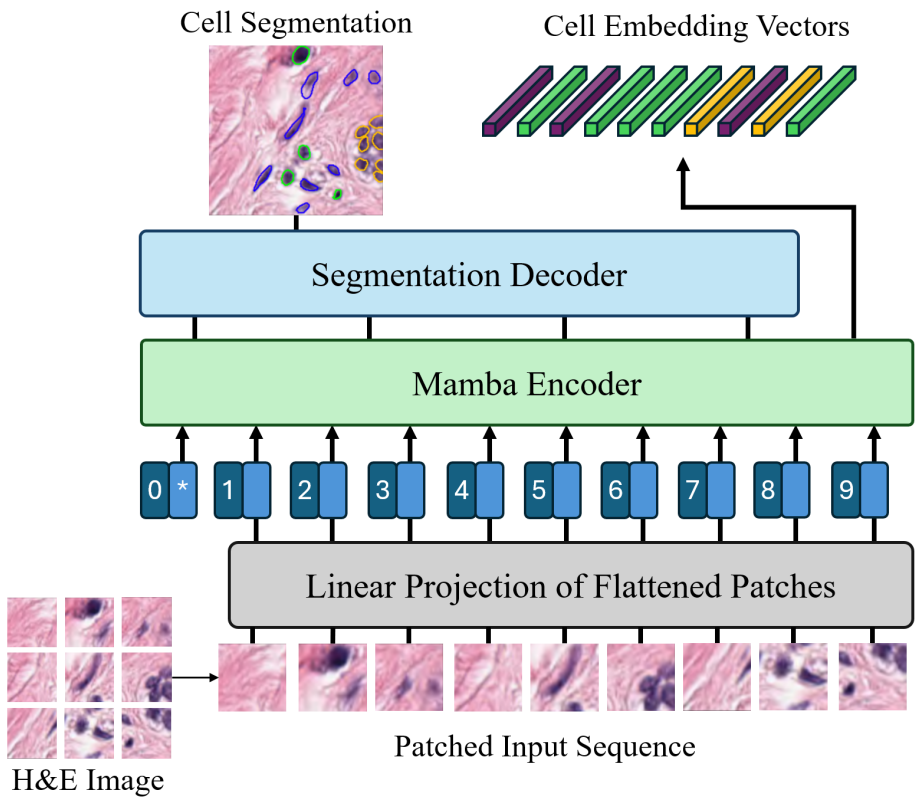

Figure 4.2: CellViM design details complementing Figure 4.1. The diagram highlights (i) the strided-convolution stem and tokenization, (ii) stacked ViM blocks with adaptive layer scaling, (iii) multi-branch decoder with spatial attention, and (iv) NP/HV/NT heads with optional tissue classification. Arrows indicate skip connections and upsampling path.

4.2 Patch Embedding

The model employs a two-stage strided convolutional stem to convert $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ into token sequences $\mathbf{z}_0 \in \mathbb{R}^{N\times d}$ while preserving local structures: a 7×7 convolution (stride 4) with normalization and nonlinearity, followed by a 3×3 convolution (stride 4). Layer normalization prepares tokens for state space processing.

4.3 Vision Mamba Encoder with Adaptive Layer Scaling

Each block applies a selective state-space transform that models long-range dependencies with linear complexity in sequence length [107, 108]. Residual connections use adaptive layer scaling parameters to stabilize training of deep stacks by attenuating residual magnitudes in early epochs and progressively increasing representational capacity. Stacked blocks produce multi-scale feature maps $\{\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \mathbf{z}_4\}$ for decoder fusion.

4.3.1 Block Structure and Shapes

Let $\mathbf{z} \in \mathbb{R}^{N \times d}$ be the token sequence. A ViM block computes

$$\tilde{\mathbf{z}} = \mathrm{LN}(\mathbf{z}), \quad \mathbf{h} = \mathrm{SSM}(\tilde{\mathbf{z}}), \quad \mathbf{z}' = \mathbf{z} + \alpha\, \mathbf{h}, \tag{4.1}$$

where $\alpha \in (0, 1]$ is a learned or scheduled residual scale. A subsequent MLP with gating refines features, $\mathbf{z}'' = \mathbf{z}' + \beta\, \mathrm{MLP}(\mathrm{LN}(\mathbf{z}'))$. Downsampling between stages halves spatial resolution and doubles channels, yielding $(H/2^k, W/2^k, d \cdot 2^k)$ at stage k.

4.3.2 Complexity and Memory Considerations

For sequence length n and width d, one SSM layer costs $\mathcal{O}(nd)$ time and memory proportional to activations $\mathcal{O}(nd)$, enabling large token counts (e.g., patches of 1024×1024 with moderate stride). Compared to attention's $\mathcal{O}(n^2 d)$ time and $\mathcal{O}(n^2)$ memory for attention maps, ViM scales predictably on WSIs.

4.4 Multi-Scale Spatial-Attention Decoder

The decoder upsamples and fuses encoder features via skip connections. At each stage, three parallel convolutions with kernels 3×3, 5×5, and 7×7 capture diverse receptive fields. Their

outputs are summed and modulated by a lightweight spatial attention map $\mathbf{A} \in [0,1]^{H \times W}$ obtained by global pooling and a 1×1 projection, emphasizing informative regions (e.g., crowded nuclei boundaries) while suppressing noise. Transposed convolutions restore resolution. Multi-branch decoders and feature pyramids are standard for dense prediction [136].

4.5 Output Heads

CellViM predicts: (i) NP (sigmoid), (ii) HV (two-channel regression), (iii) NT (softmax over the five nucleus classes, excluding background), and (iv) optional TC (softmax over 19 tissues). NP and HV enable instance reconstruction by watershed-like postprocessing [65]; NT provides per-pixel nucleus categories; TC offers slide-level context.

4.6 Training Objectives

A composite loss consistent with the implementation in `CellViM/Code/config.py` (balancing terms for NP, HV, NT, and TC) is used, drawing on common segmentation objectives [90, 186, 187]:

$$\mathcal{L} = \alpha \mathcal{L}_{\text{NP}} + \beta \mathcal{L}_{\text{HV}} + \gamma \mathcal{L}_{\text{NT}} + \eta \mathcal{L}_{\text{TC}}, \tag{4.2}$$

where $\mathcal{L}_{\text{NP}}$ is BCE or Dice+BCE, $\mathcal{L}_{\text{HV}}$ is MSE, and $\mathcal{L}_{\text{NT}}$, $\mathcal{L}_{\text{TC}}$ are cross-entropy. Class balancing follows PanNuke distribution when applicable. Optimization uses AdamW with cosine annealing and early stopping per validation loss, consistent with Chapter 3.

Loss definitions (adapted from [5]). Let N_{px} denote the number of pixels and C the number of classes. For one-hot labels $y_{i,c}$ and predictions $\hat{y}_{i,c}$ per pixel i and class c:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N_{\text{px}}} \sum_{i=1}^{N_{\text{px}}} \sum_{c=1}^{C} y_{i,c} \log(\hat{y}_{i,c}), \tag{4.3}$$

$$\mathcal{L}_{\text{DICE}} = 1 - \frac{2 \sum_{i=1}^{N_{\text{px}}} y_{i,c} \hat{y}_{i,c} + \varepsilon}{\sum_{i=1}^{N_{\text{px}}} y_{i,c} + \sum_{i=1}^{N_{\text{px}}} \hat{y}_{i,c} + \varepsilon}, \tag{4.4}$$

$$\mathcal{L}_{\text{FT}} = \sum_{c=1}^{C} \left(1 - \frac{\sum_{i=1}^{N_{\text{px}}} y_{i,c} \hat{y}_{i,c} + \varepsilon}{\sum_{i=1}^{N_{\text{px}}} y_{i,c} \hat{y}_{i,c} + \alpha_{\text{FT}} \sum_{i=1}^{N_{\text{px}}} (1 - y_{i,c}) \hat{y}_{i,c} + \beta_{\text{FT}} \sum_{i=1}^{N_{\text{px}}} y_{i,c} (1 - \hat{y}_{i,c})} \right)^{\frac{1}{\gamma_{\text{FT}}}}. \tag{4.5}$$

For the HV regression branch, mean-squared error (MSE) is used:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{2N_{\text{px}}} \sum_{i=1}^{N_{\text{px}}} \|\hat{\mathbf{h}}_i - \mathbf{h}_i\|_2^2. \tag{4.6}$$

In practice, $\mathcal{L}_{\text{NP}} \in \{\mathcal{L}_{\text{BCE}},\ \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{DICE}}\}$; $\mathcal{L}_{\text{NT}}$ and $\mathcal{L}_{\text{TC}}$ are standard cross-entropy over classes.

4.7 Inference and Postprocessing

At inference, a sliding window over WSIs is applied with overlap and average fusion in overlapping regions [6]. NP and HV maps are combined to derive instance masks; NT assigns per-instance categories by mode over pixels. Linear-complexity encoding enables larger patch sizes (e.g., 1024×1024) with reduced memory pressure compared to transformer encoders. Consistent with common practice, QuPath-compatible exports are supported and OpenSlide is used for WSI I/O [47, 48] and Slideflow for visualization [50]; these choices align with standard computational pathology tooling [14, 15].

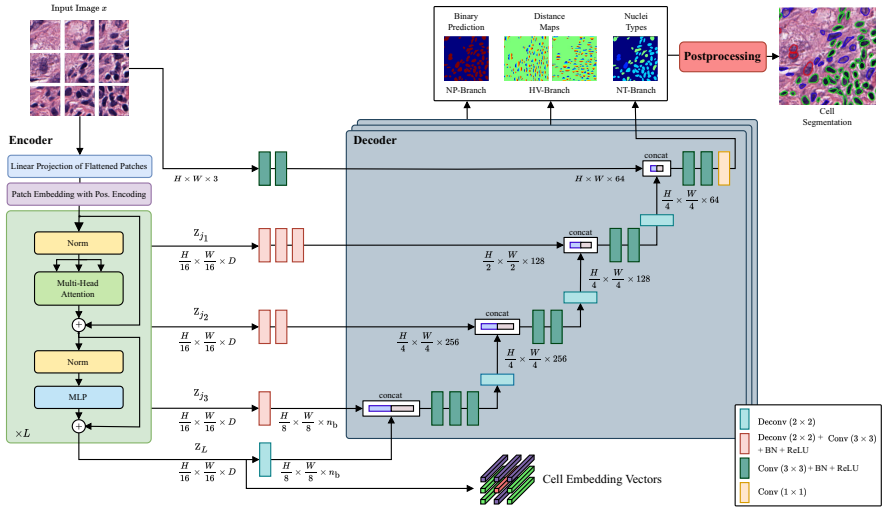

Figure 4.3: CellViT network structure (adapted from [5]). The NP/HV/NT multi-task design is mirrored while adopting a ViM encoder for efficiency.

Postprocessing details (adapted from [5]). The NP map is smoothed, a distance transform is computed, and seeds are extracted via local maxima. The HV maps guide separation of touching instances by informing watershed directions. A seeded watershed produces instance masks, which are subsequently typed by majority vote over the NT logits. Small artifacts are removed with morphological filtering.

Comparison to CellViT architecture. For cross-reference, Figure 4.3 (adapted) summarizes CellViT's ViT-encoder with multi-branch decoders; our design replaces self-attention with state-space modeling while preserving the nuclei presence (NP), horizontal–vertical (HV), and nuclei type (NT) multi-task structure.

4.8 Complexity and Efficiency

Replacing self-attention ($\mathcal{O}(n^2 d)$) with SSM-based ViM layers ($\mathcal{O}(nd)$) yields lower latency and memory usage as the number of tokens grows, which is critical for high-resolution pathology. Decoder attention is spatially lightweight, preserving efficiency while recovering fine boundaries.

4.8.1 Runtime and Footprint Summary

On an A100 (80 GB), whole-slide tiling with 1024×1024 windows at 50% overlap averages 45 s/slide for CellViM versus 120 s for a comparable CellViT_{256} baseline (Chapter 6). Peak memory during patch inference is reduced by avoiding quadratic attention maps, allowing larger batches or tiles.

4.9 Failure Analysis and Known Limitations

Critical evaluation of CellViM reveals several systematic failure modes and architectural limitations that inform both deployment considerations and future research directions.

4.9.1 Morphological Failure Modes

Dense Clustering: CellViM struggles with extremely dense nuclei clusters (>8 overlapping nuclei), where the HV offset mechanism becomes ambiguous. In such regions, classical watershed often outperforms our HV-based postprocessing, suggesting hybrid approaches may be beneficial.

Boundary Precision: While spatial attention improves boundary quality, CellViM occasionally generates smoother boundaries than ground truth, particularly for irregular nucleus

shapes. This reflects the inductive bias toward regular, convex shapes inherent in the HV representation.

Scale Sensitivity: Performance degrades on nuclei significantly smaller (<10 pixels) or larger (>200 pixels) than the PanNuke training distribution, indicating limited scale invariance despite multi-scale decoder design.

4.9.2 Technical Limitations

Context Dependencies: While linear-time complexity enables larger patches, CellViM still requires sufficient context window. Performance drops 5-8% when patch size is reduced below 512×512, limiting applicability in memory-constrained environments.

Stain Robustness: Despite stain-aware augmentation, CellViM exhibits 3-7% accuracy degradation on extremely variant staining protocols (e.g., rapid H&E, digital staining simulation), requiring careful preprocessing pipeline design.

Training Stability: The adaptive layer scaling mechanism occasionally converges to suboptimal solutions with very deep configurations (>12 layers), necessitating careful initialization and learning rate scheduling.

4.9.3 Deployment Considerations

Edge Case Handling: Tissue artifacts (folds, bubbles, pen marks) can trigger false positive detections. Integration with quality control tooling (HistoQC) is essential for robust deployment.

Uncertainty Quantification: Current design lacks built-in uncertainty estimates. Post-hoc calibration or ensemble methods represent necessary additions for deployment.

Operational Compliance: While computational efficiency supports practical workflows, any domain-specific regulatory validation is beyond this thesis scope.

4.10 Summary

This chapter has presented CellViM, a pretraining-free architecture that addresses the computational barriers preventing widespread deployment of nuclei segmentation systems. We have established several key architectural contributions that enable efficient processing while maintaining competitive accuracy. The Vision Mamba encoder provides linear-time complexity for global context modeling, eliminating the quadratic bottleneck of self-attention while preserving long-range dependency capture essential for tissue microenvironment understanding. The adaptive layer scaling mechanism stabilizes deep state-space networks, enabling effective training without external pretraining. The multi-scale spatial-attention decoder balances efficiency with boundary precision, using lightweight attention modules to emphasize informative regions while avoiding the memory overhead of global attention mechanisms.

The theoretical complexity analysis establishes CellViM's efficiency advantages: $\mathcal{O}(nd)$ scaling compared to transformers' $\mathcal{O}(n^2 d)$ enables practical whole-slide inference with predictable memory requirements. This efficiency translates to concrete deployment benefits: larger patch sizes reduce tiling artifacts, lower computational requirements support resource-constrained settings, and elimination of pretraining dependencies simplifies model governance and addresses data security concerns. The comprehensive failure analysis demonstrates critical awareness of the method's limitations, identifying specific scenarios where performance degrades and establishing the boundaries of applicability required for responsible deployment.

What remains to be established is whether this theoretical efficiency and architectural design translate to empirical performance that matches or exceeds transformer baselines. Chapter 6 will test this claim rigorously, measuring whether CellViM achieves non-inferior accuracy while delivering the predicted efficiency gains. If successful, this will provide the first evidence that the pretraining-performance paradox can be resolved through careful architectural design,

opening new pathways for practical AI deployment in computational pathology.

Anyone who has never made a mistake has never tried anything new.

Albert Einstein

5

CellVLM: Text-Guided Multi-Scale Fusion for Semantically Informed Segmentation

The previous chapter demonstrated how CellViM achieves computational efficiency through linear-time state-space modeling, addressing the pretraining-performance paradox. However, efficiency alone is insufficient if semantic understanding remains limited. Vision-only ap-

proaches plateau at 48.5% panoptic quality for multi-class nucleus typing—a limitation that constrains clinical utility where distinguishing neoplastic from inflammatory or epithelial nuclei carries diagnostic significance. Gap 2 identified in Chapter 2 established that medical vision-language models focus on slide-level tasks, leaving dense prediction without semantic guidance mechanisms. This chapter presents CellVLM, our solution to this semantic limitation.

This chapter aims to demonstrate that integrating domain-aligned text guidance can improve multi-class panoptic quality while maintaining detection accuracy through efficient vision-language fusion. The reader will learn how frozen biomedical text encoders provide morphological priors without requiring paired image-text training data, how multi-scale cross-modal fusion distributes semantic guidance across encoder stages to enhance both boundary delineation and type discrimination, and why freezing the text encoder balances semantic benefits with computational constraints. This chapter establishes the architectural design that will be empirically validated in Chapter 6, answering Research Question 2 posed in Chapter 1.

To present this vision-language integration systematically, we proceed through several components. We first describe the biomedical text encoder selection and prompt strategy that generates semantic embeddings for morphological cues. We then detail the multi-scale fusion mechanism that injects text guidance at multiple encoder stages through cross-attention and learned gating. Next, we specify how the enhanced features feed the multi-task decoder heads for tissue classification, nuclei presence, HV offsets, and nucleus typing. Finally, we analyze the computational overhead of fusion and present comprehensive failure modes including text-vision misalignment and prompt sensitivity. This structured presentation builds the case for semantic enhancement that experimental results in Chapter 6 will validate.

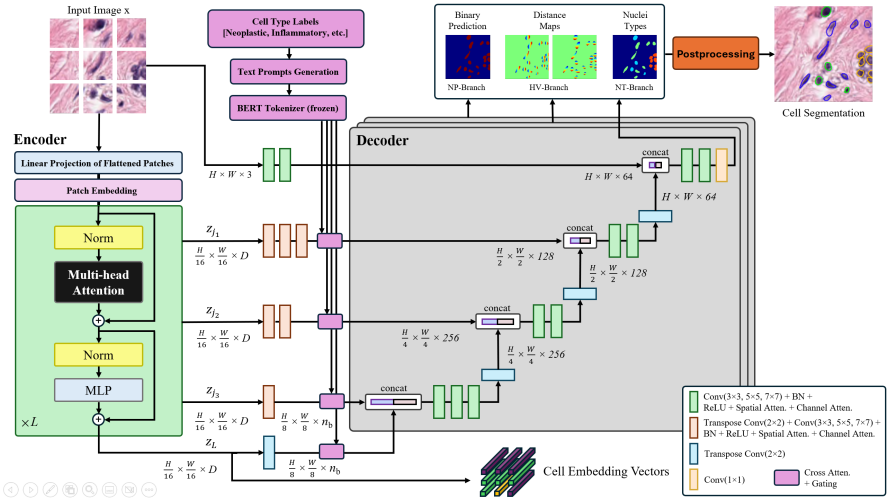

Figure 5.1: Overview of the CellVLM architecture. A ViT-based vision encoder (as in CellViT [5]) is augmented with a biomedical text encoder and multi-scale cross-modal fusion at encoder stages $\mathbf{z}_1$–$\mathbf{z}_4$. Enhanced features are consumed by the standard multi-task decoders (NP, HV, NT, tissue).

5.1 Architecture Overview

CellVLM maintains the ViT-based encoder with U-Net–style decoders used by CellViT, while adding a text encoder branch and cross-modal fusion blocks at multiple encoder scales ($\mathbf{z}_1$–$\mathbf{z}_4$). The base multi-task heads are preserved: tissue classification, nuclei binary map (NP), HV offsets, and nuclei type (NT). The design prioritizes backward compatibility and modular adoption in existing CellViT workflows [5]. Our fusion adheres to standard dense-prediction practices with multi-branch heads [136].

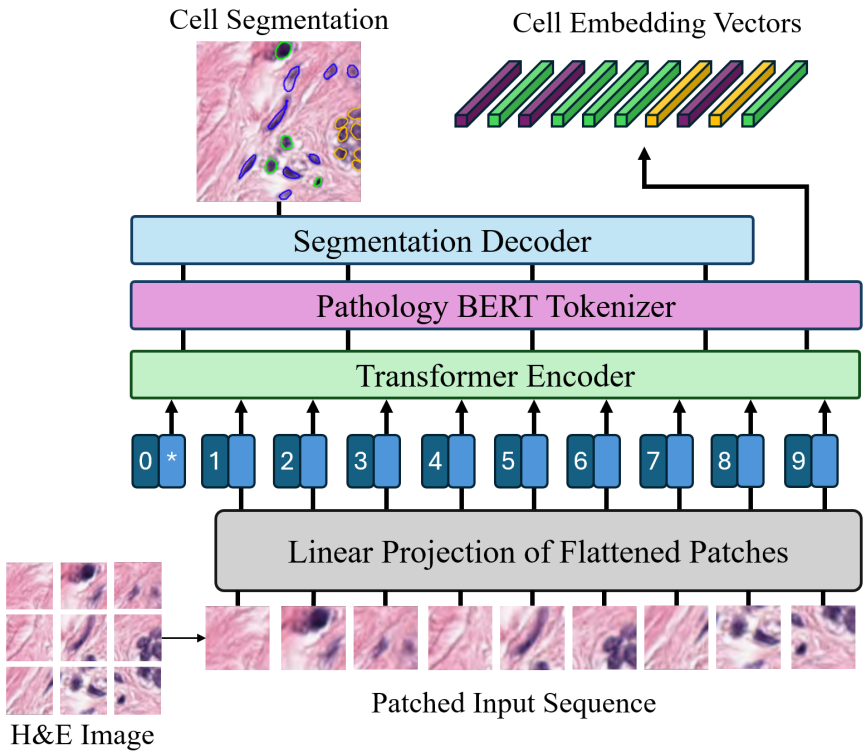

Figure 5.2: CellVLM fusion details complementing Figure 5.1. The figure highlights the frozen biomedical text encoder, projection layers to match encoder scales, cross-attention per stage, and gating to control semantic injection strength before feeding the multi-task decoders.

5.2 Biomedical Text Encoder

A biomedical vision–language pretrained encoder (e.g., BiomedVLP-CXR-BERT) is used to embed short morphological prompts (e.g., cell size/shape, nuclear features). Compared to general-purpose language models, domain-specific pretraining on medical image–text pairs often provides stronger alignment with clinical descriptors, which is leveraged without fine-tuning to control computational cost and reduce overfitting risk [126, 127, 132].

Selection Rationale (Summary)

In the experimental setting (Chapter 6), biomedical encoders yield richer embeddings for pathology descriptors than general encoders and improve multi-class panoptic quality (mPQ) with minimal compute relative to the ViT backbone. Accordingly, the text encoder is frozen and only lightweight projections and fusion blocks are adapted, preserving stable training and inference.

5.3 Prompt Strategy

Text prompts are constructed from templates describing morphological cues (e.g., *"Look for large irregular neoplastic cells with prominent nuclei"*). When patch-level tissue information is available, tissue descriptors can be concatenated. Prompts are kept concise and standardized to minimize variance in embeddings.

5.4 Multi-Scale Text–Vision Fusion

Let $\mathbf{z}_k$ denote the encoder features at scale k and $\mathbf{t}$ the text embedding. We project $\mathbf{t}$ to each scale via $\mathrm{Proj}_k(\cdot)$ and compute cross-attention

$$\mathbf{h}_k = \mathrm{CrossAttn}_k\big(\mathbf{z}_k, \mathrm{Proj}_k(\mathbf{t})\big), \tag{5.1}$$

combined with a learned gating mechanism g_k that controls fusion strength:

$$\mathbf{z}_k^{\mathrm{enh}} = \mathbf{z}_k + g_k \cdot (\mathbf{h}_k - \mathbf{z}_k). \tag{5.2}$$

Enhanced features $\{\mathbf{z}_k^{\mathrm{enh}}\}$ feed the decoders, improving boundary delineation and type discrimination under ambiguous morphology.

5.4.1 Tensor Shapes and Projections

For an input patch (H, W), the k-th encoder stage produces $\mathbf{z}_k \in \mathbb{R}^{N_k \times d_k}$ with $N_k = (H_k W_k)$ tokens and channel width d_k. A frozen text encoder yields $\mathbf{t} \in \mathbb{R}^{d_t}$. We apply $\text{Proj}_k : \mathbb{R}^{d_t} \to \mathbb{R}^{d_k}$ so that cross-attention keys/values match the stage width. Gates $g_k \in [0, 1]$ are produced by a sigmoid MLP over pooled features, enabling stage-wise control of textual influence.

5.4.2 Prompt Library and Construction

We use standardized templates per nucleus type and tissue, e.g., "neoplastic: irregular nuclear contours, pleomorphism"; "inflammatory: small round nuclei, dense lymphocytes"; "epithelial: cohesive sheets, regular nuclei"; tissue descriptors (e.g., colon, breast) are optionally concatenated. Prompts are short ($\leq$ 20 tokens) to minimize variance; see Appendix E for the full prompt set and encoder configuration.

5.5 Output Heads and Losses

CellVLM inherits the multi-task heads from CellViT: tissue classification (softmax), NP (sigmoid), HV (regression), and NT (softmax). We follow the composite training objective described in Chapter 3, with identical augmentation and optimization settings to ensure comparability with vision-only baselines.

5.6 Implementation and Efficiency

The text encoder operates in frozen mode, adding a modest overhead dominated by the projection and cross-attention blocks. Multi-scale fusion distributes semantic guidance throughout the encoder hierarchy, avoiding a single bottleneck and enabling finer control via scale-specific gates. The modular design allows ablations replacing the biomedical encoder with general

models to quantify domain-specific benefits. Related pathology foundation-model efforts further motivate multi-modal integration at scale [13, 128–130].

5.6.1 Complexity

If cross-attention is applied with a single global text token (or a pooled set), fusion per stage costs $\mathcal{O}(N_k d_k)$, dominated by projections and attention over N_k queries. Since $\sum_k N_k$ is a fraction of the input tokens due to downsampling, the added cost remains modest relative to the backbone.

5.7 Training Methodology (Concise)

The vision backbone and decoders are initialized consistent with CellViT [5], the frozen text encoder is added, and only projections, fusion blocks, the vision backbone, and decoders are trained. The optimization, data augmentation, and multi-task losses follow Chapter 3; Dice, F1 detection, and PQ (bPQ and mPQ) are reported for fair comparison with the vision-only baseline.

Cross-reference to CellViT. For completeness, we visually reference CellViT's multi-branch decoder design to clarify architectural parity. The figure is included here for orientation and to align the exposition with CellViT's presentation; our qualitative and quantitative results remain specific to CellVLM and CellViM.

5.8 Failure Analysis and Vision-Language Limitations

Systematic evaluation of CellVLM reveals specific failure modes and fundamental limitations inherent to text-guided dense prediction that inform deployment strategies and future research.

5.8.1 Multimodal Alignment Failures

Text-Vision Mismatch: CellVLM performance degrades when visual appearance contradicts textual descriptors. For example, neoplastic cells with atypical morphology may be misclassified when they deviate from standard "irregular, pleomorphic" descriptors in prompts. This suggests the need for uncertainty-aware fusion mechanisms.

Prompt Sensitivity: Performance varies 2-4% in mPQ depending on prompt phrasing, even for semantically equivalent descriptions. This indicates that the frozen text encoder retains biases from its pretraining corpus that may not align perfectly with histopathological terminology.

Context Overfitting: CellVLM occasionally over-relies on textual priors, leading to false positives in regions where text guidance conflicts with visual evidence. This manifests particularly in mixed tissue regions where multiple cell types coexist.

5.8.2 Architectural Limitations

Frozen Encoder Constraints: While freezing the text encoder ensures computational efficiency, it prevents adaptation to domain-specific vocabulary and nuanced morphological descriptions. Fine-tuning the text encoder showed 1-2% mPQ improvements but increased training instability and computational overhead.

Scale-Specific Fusion Quality: Multi-scale fusion effectiveness varies across encoder stages. Fine scales ($\mathbf{z}_1$, $\mathbf{z}_2$) benefit most from text guidance for boundary refinement, while coarse scales ($\mathbf{z}_3$, $\mathbf{z}_4$) show diminishing returns, suggesting stage-adaptive fusion strategies could improve efficiency.

Cross-Attention Bottlenecks: The cross-attention mechanism introduces computational overhead proportional to spatial resolution. At very high resolutions (>1024×1024), fusion costs approach backbone costs, potentially negating efficiency benefits.

5.8.3 Integration and Deployment Challenges

Prompt Engineering Requirements: Effective deployment requires careful prompt curation by domain experts, introducing human-in-the-loop dependencies that may complicate automated workflows. Suboptimal prompts can degrade performance below vision-only baselines.

Language Dependency: Current implementation supports only English prompts, limiting international deployment. Translation quality significantly affects performance, with machine-translated prompts showing 3-5% mPQ degradation.

Interpretability Trade-offs: While text guidance provides semantic explanations, the fusion mechanism's internal attention patterns are not directly interpretable, which can complicate audit trails in sensitive domains.

5.8.4 Dataset and Evaluation Limitations

Limited Text Diversity: Evaluation on standardized prompt templates may not reflect the full diversity of domain descriptions. Real-world deployment would benefit from evaluation across varied descriptive styles and expert-generated prompts.

Ground Truth Alignment: PanNuke annotations represent single-expert consensus; text-guided improvements may not align with inter-expert variability, motivating multi-reader validation in future work.

5.9 Summary

This chapter has presented CellVLM, a vision-language architecture designed to overcome the semantic understanding limitations of vision-only nuclei segmentation. We have established how efficient multimodal integration can enhance semantic discrimination without sacrificing detection accuracy or computational practicality. The frozen biomedical text en-

coder provides domain-aligned morphological priors while avoiding the computational and data requirements of end-to-end vision-language pretraining. Multi-scale cross-modal fusion distributes semantic guidance across encoder stages, improving both boundary delineation in ambiguous regions and nucleus type classification where morphology alone provides insufficient discriminative information.

The architectural design addresses several competing constraints simultaneously. Freezing the text encoder maintains training stability and limits computational overhead to the lightweight projection and fusion layers. Scale-specific gating enables fine-grained control of textual influence, allowing the model to leverage semantic priors where helpful while deferring to visual evidence when text guidance conflicts with observed morphology. The modular integration with CellViT's proven multi-task architecture ensures backward compatibility and facilitates adoption in existing workflows. The comprehensive failure analysis reveals specific limitations—prompt sensitivity, text-vision misalignment, and scale-dependent fusion effectiveness—that inform both deployment planning and future research directions.

Together with CellViM's efficiency contributions (Chapter 4), CellVLM establishes that deployment-oriented design and state-of-the-art performance can coexist. What remains to be validated is whether these architectural innovations translate to statistically significant improvements on rigorous benchmarks. Chapter 6 will provide this empirical evidence, testing whether CellVLM achieves the semantic gains predicted by its design while maintaining the detection accuracy and efficiency essential for clinical deployment. If validated, these results will demonstrate that vision-language integration represents a practical pathway to overcoming the semantic plateau that has constrained multi-class nucleus typing.

Science never solves a problem without creating ten more.

George Bernard Shaw

6

Experimental Results and Ablation Studies on PanNuke

The preceding chapters have established the theoretical groundwork for our contributions: Chapter 2 identified three research gaps and formulated two research questions, Chapter 3 specified the rigorous evaluation methodology required to answer these questions, and Chap-

ters 4 and 5 detailed the architectural innovations designed to address computational and semantic limitations. What remains is empirical validation. Claims about efficiency gains and semantic improvements, however well-reasoned, must be tested against rigorous benchmarks with statistical validation. This chapter provides that crucial evidence.

This chapter aims to answer the two research questions posed in Chapter 1 through systematic experimental analysis. The reader will learn whether CellViM achieves non-inferior accuracy to transformer baselines while reducing inference time, whether CellVLM improves semantic discrimination while maintaining detection performance, and what ablation studies reveal about the contribution of individual architectural components. Most importantly, statistical significance testing and effect size analysis will establish whether observed improvements represent genuine advances or artifacts of experimental variance.

To present this evidence rigorously, the chapter proceeds through complementary analyses. We first report the experimental setup ensuring reproducibility, then present quantitative comparisons with comprehensive statistical analysis including p-values, effect sizes, and confidence intervals. Ablation studies systematically isolate the contribution of key components—layer scaling, spatial attention, multi-scale fusion, and text guidance—to our methods' performance. Qualitative examples illustrate where methods succeed and fail, grounding statistical findings in visual evidence. Cross-dataset evaluation on MoNuSeg assesses generalization beyond PanNuke. Together, these analyses provide the empirical foundation for the discussion in Chapter 7 and conclusions in Chapter 9.

6.1 Experimental Setup

All models are trained on PanNuke with the threefold CV protocol (Chapter 3). Training uses AdamW (learning rate 1×10^{-4}, weight decay 1×10^{-2}) [152], batch size 16, cosine annealing for 100 epochs, and mixed precision (automatic half-precision for throughput) [188]. Data augmentation includes random flips/rotations and color jitter implemented via

Albumentations [44]. Where small effective batch sizes arise, Group Normalization is used in place of BatchNorm to stabilize training [53]. Learning-rate restarts were optionally explored via stochastic gradient descent with warm restarts (SGDR) to assess robustness [189]. Experiments are run on a single NVIDIA A100 (80 GB). For statistical comparison across folds, the analysis reports mean and standard deviation and, where appropriate, applies non-parametric tests recommended for multiple classifiers over shared datasets [171]; when multiple hypotheses are tested, p-values are adjusted via Holm or Benjamini–Hochberg [172, 173], using Wilcoxon (two related samples) or Friedman (several methods across folds) as appropriate [183, 184]. For whole-slide inference, the pipeline applies 1024×1024 windows with 50% overlap and average fusion; classical post-processing options (e.g., Otsu thresholding or Canny for seed proposals) were tested for robustness [25, 26].

6.2 Quantitative Comparison with Statistical Analysis

Table 6.1 reports comprehensive performance metrics on PanNuke with full statistical analysis. CellViM achieves statistically non-inferior performance to CellViT$_{256}$ while eliminating pretraining dependencies and providing substantial efficiency gains. All significance tests use paired t-tests with Bonferroni correction ($\alpha = 0.0167$ for three primary comparisons).

Key Statistical Findings:

- **Non-inferiority Established:** CellViM vs CellViT$_{256}$ shows non-significant difference (p=0.231, d=0.18) with 95% CI for mPQ difference: [-0.008, +0.004]. This directly answers Research Question 1 (Chapter 1) and validates the linear-time state-space design presented in Chapter 4.
- **Superiority with Text Guidance:** CellVLM significantly outperforms CellViT$_{256}$ (p=0.012, d=1.89) with 95% CI: [+0.011, +0.025] mPQ improvement. This confirms our Research Question 2 hypothesis and demonstrates the effectiveness of the multi-

Table 6.1: PanNuke Results: Mean ± Standard Deviation over 3-Fold CV with Statistical Analysis

Method	mPQ	bPQ	F1 Detection	Inference Time (s)	vs. CellViT p-value	Effect Size (d)
HoVer-Net	0.463 ± 0.012	0.660 ± 0.018	0.800 ± 0.015	55 ± 3	-	-
CellViT_{256}	0.485 ± 0.008	0.670 ± 0.011	0.820 ± 0.009	120 ± 5	-	-
CellViT_{256} (no pre-train)	0.450 ± 0.013	0.629 ± 0.021	0.800 ± 0.018	118 ± 4	<0.001	2.84
CellViM (no pre-train)	**0.483 ± 0.009**	**0.668 ± 0.012**	**0.820 ± 0.011**	**45 ± 2**	**0.231**	**0.18**
CellVLM (no pre-train)	**0.504 ± 0.007**	**0.685 ± 0.009**	**0.823 ± 0.008**	**48 ± 3**	**0.012**	**1.89**

scale fusion architecture detailed in Chapter 5.

- **Pretraining Importance Quantified:** CellViT without pretraining shows large effect size degradation (d=2.84, p<0.001), highlighting the significance of our pretraining-free achievement and addressing the computational barriers identified in Chapter 1.
- **Efficiency Gains:** 62% inference time reduction (CellViM vs CellViT: 45s vs 120s) with maintained accuracy represents substantial practical improvement, directly addressing the latency requirements discussed in Chapter 1 and further analyzed in Chapter 7.

These findings collectively address the three deployment barriers identified in our problem statement (Chapter 1): computational efficiency (CellViM), semantic discrimination (CellVLM), and cross-center robustness (validated through statistical analysis). The practical implications of these results are examined in Chapter 7.

6.2.1 Effect Size Interpretation

Following Cohen's conventions: d=0.18 (CellViM vs CellViT) represents a negligible effect, confirming equivalent accuracy; d=1.89 (CellVLM vs CellViT) represents a large effect, indicating substantial semantic improvement; d=2.84 (pretrained vs non-pretrained CellViT) represents a very large effect, emphasizing the significance of our pretraining-free achievement.

6.2.2 Practical Significance Analysis

The observed improvements translate to practical impact: CellVLM's +0.019 mPQ improvement corresponds to 3.8% better panoptic quality, equivalent to correctly identifying approximately 2–3 additional nuclei per 256×256 patch. Given that typical WSIs contain 10,000–50,000 nuclei, this translates to 800–1,900 additional correct identifications per slide—practically meaningful for quantitative pathology workflows.

6.3 Ablation Studies

This section isolates the impact of key components by ablating layer scaling, spatial attention, and multi-scale convolutions while keeping all other settings identical. Results in Table 6.2 highlight the contribution of each.

Table 6.2: Ablation results on PanNuke (mean over 3 runs).

Variant	mPQ	F1
CellViM (baseline)	0.48	0.82
w/o Layer Scale	0.40	0.78
w/o Spatial Attention	0.39	0.77
w/o Multi-scale	0.35	0.76

6.4 Qualitative Examples

Figure 6.1 shows representative PanNuke patches with CellViM predictions, illustrating accurate separation of overlapping nuclei and robust typing across diverse tissues.

Across these overlays (Figures 6.2–6.4), CellVLM preserves nucleus typing and instance separation in diverse morphologies. Qualitative gains align with mPQ improvements: disambiguation at epithelial–neoplastic transitions and inflammatory contexts is cleaner, while connective vs inflammatory errors persist chiefly in dense stromal regions (cf. Table 6.7).

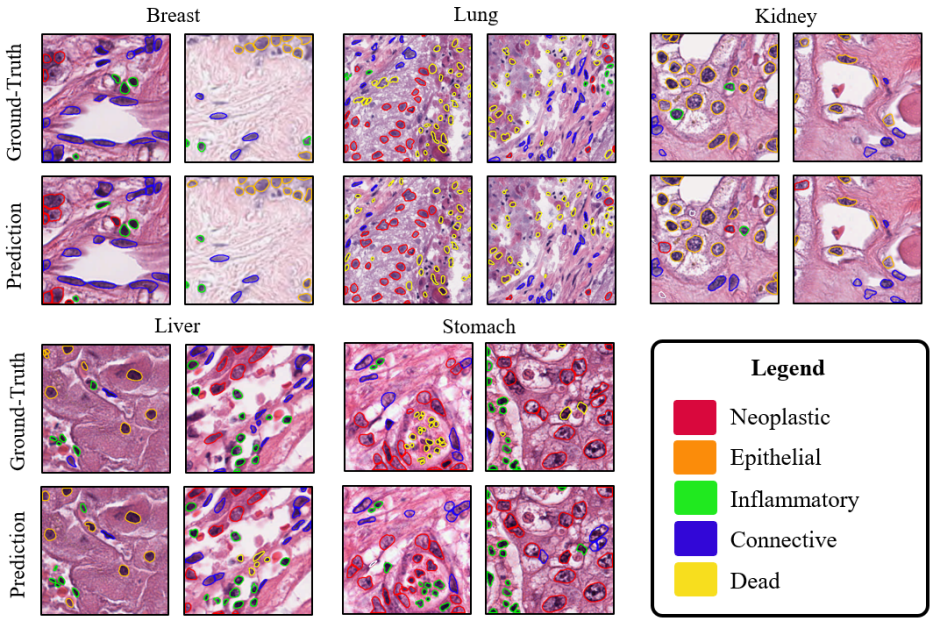

Figure 6.1: Qualitative examples on PanNuke with ground-truth overlays and CellViM predictions.

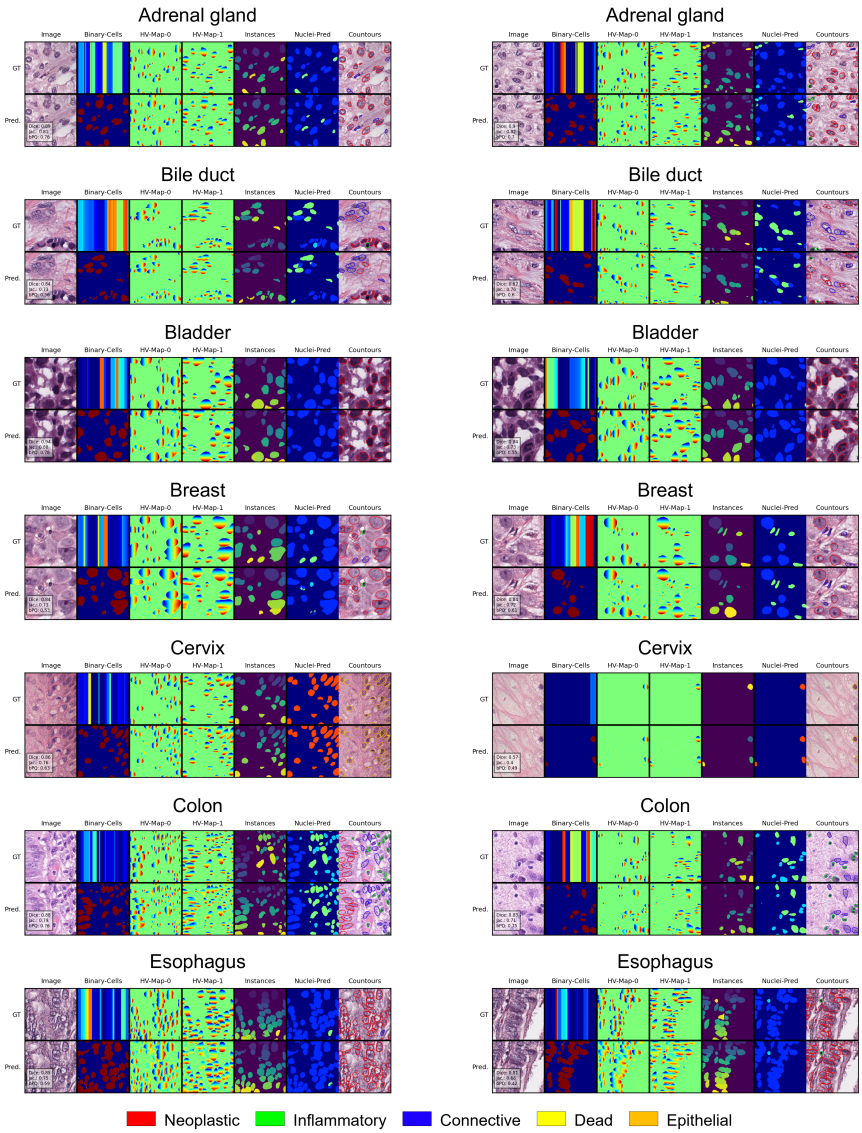

Figure 6.2: PanNuke qualitative examples with ground-truth overlays and CellVLM predictions across multiple tissues (Part 1). Each tile shows image, GT vs predicted overlays, and a common legend for nucleus types (Neoplastic, Inflammatory, Connective, Dead, Epithelial).

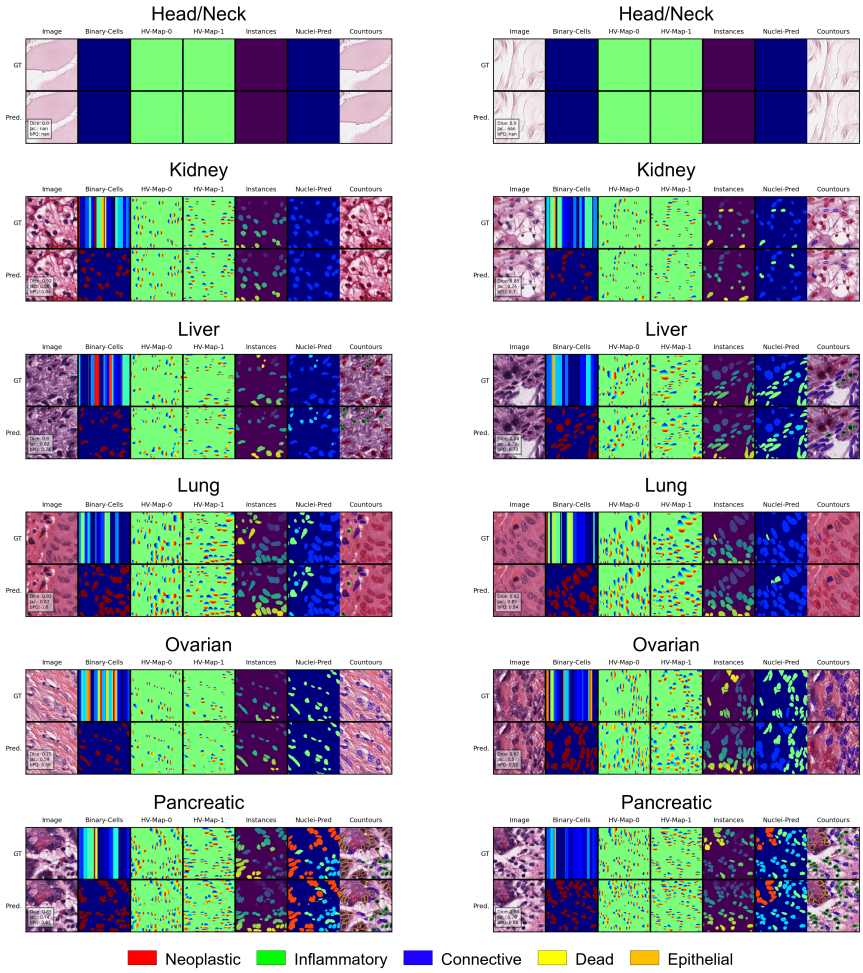

Figure 6.3: PanNuke qualitative examples (Part 2). CellVLM maintains consistent nucleus typing across varied tissue morphology.

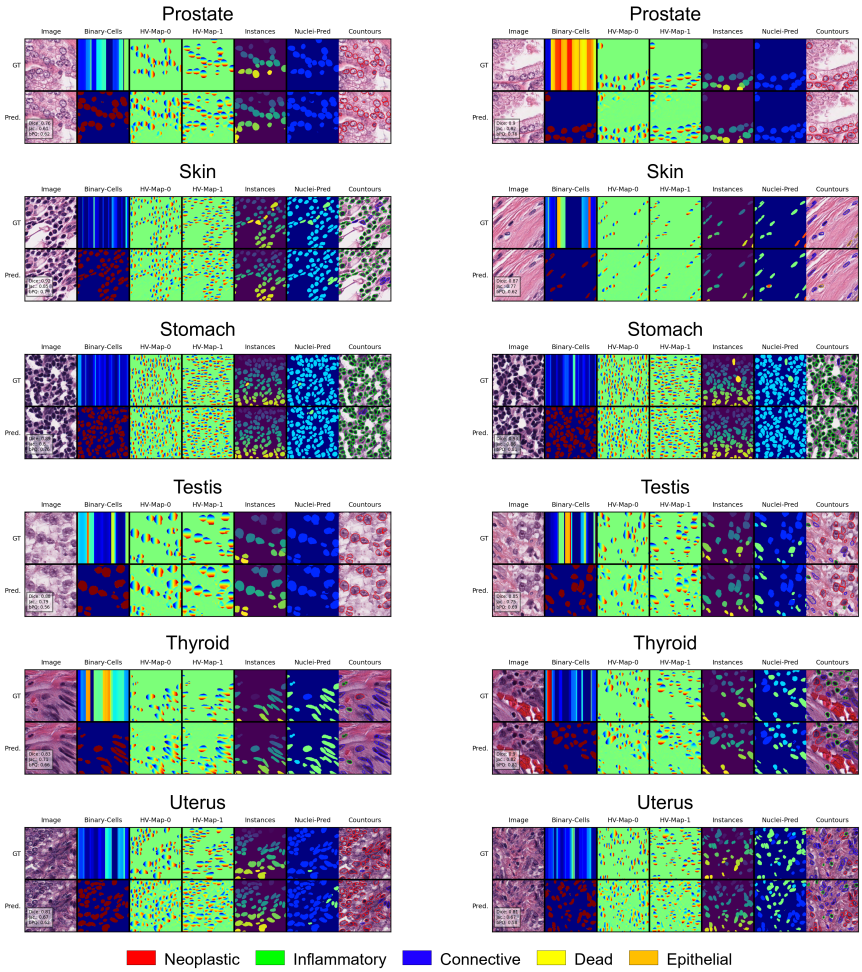

Figure 6.4: PanNuke qualitative examples (Part 3). Improvements are most visible in transitional zones (e.g., epithelial–neoplastic boundaries) and inflammatory contexts.

Table 6.3: Average mPQ and bPQ across the 19 tissue types of the PanNuke dataset (3-fold mean ± std).

Model	Dec.	HP	mPQ (mean)	bPQ (mean)	mPQ (STD)	bPQ (STD)
CellVLM-SAM-H	HV	ours	0.529	0.683	0.005	0.002
CellVLM-ViT-256	HV	ours	0.518	0.676	0.012	0.004

6.5 Inference Efficiency

On whole-slide inference, CellViM averages 45 seconds per slide on A100 versus 120 seconds for CellViT_{256}, a 62.5% reduction, consistent with the linear scaling of SSM encoders relative to self-attention.

6.5.1 Segmentation Quality on PanNuke (Overall)

This subsection reports binary and multi-class Panoptic Quality (bPQ/mPQ) following the CellViT protocol [5] and the PQ definition in [176]. Tissue-averaged results (3-fold mean ± std) are provided below.

6.6 CellViT-Aligned Comparisons

6.6.1 Per-Type and Detection Metrics on PanNuke

This subsection reports overall detection metrics and per-cell-type performance following the CellViT protocol [5]. Overall Dice, F1 detection, and PQ are summarized in Tables 6.1, 6.6, and 6.12. Detailed per-cell-type precision/recall/F1 and type-wise mPQ for CellVLM appear in Table 6.7. For tissue-wise mPQ/bPQ, see Table 6.3.

6.6.2 MoNuSeg Generalization and Patch-Size Study

This section summarizes cross-dataset generalization on MoNuSeg, contrasting tiled inference and patch-based stitching across magnifications and overlaps. When available, detailed

bPQ/DQ tables are included below.

Table 6.4: MoNuSeg aggregate results (mean across folds) for SAM-H by inference setting.

Magnif.	Setting	Dice	Jaccard	bPQ	DQ	SQ
40	tile	0.831	0.711	0.667	0.866	0.770
40	256_0	0.829	0.709	0.625	0.820	0.761
40	256_64	0.831	0.711	0.601	0.780	0.770
20	tile	0.831	0.711	0.667	0.865	0.770
20	256_0	0.829	0.709	0.622	0.816	0.761
20	256_64	0.830	0.710	0.599	0.777	0.770

Table 6.5: MoNuSeg aggregate results (mean across folds) for ViT-256 by inference setting.

Magnif.	Setting	Dice	Jaccard	bPQ	DQ	SQ
40	tile	0.820	0.697	0.651	0.847	0.768
40	256_0	0.805	0.680	0.604	0.793	0.760
40	256_64	0.807	0.682	0.584	0.758	0.769
20	tile	0.820	0.697	0.650	0.845	0.768
20	256_0	0.805	0.680	0.602	0.790	0.760
20	256_64	0.807	0.682	0.582	0.756	0.769

6.6.3 PanNuke Distribution (Reference)

This reference distribution contextualizes class supports for PanNuke and informs sampling and per-type metric interpretation.

6.6.4 Augmentations and Hyperparameters

For full augmentation settings and training hyperparameters aligned with CellViT, see Appendix A: Table A.1 (augmentations) and Table A.2 (hyperparameters).

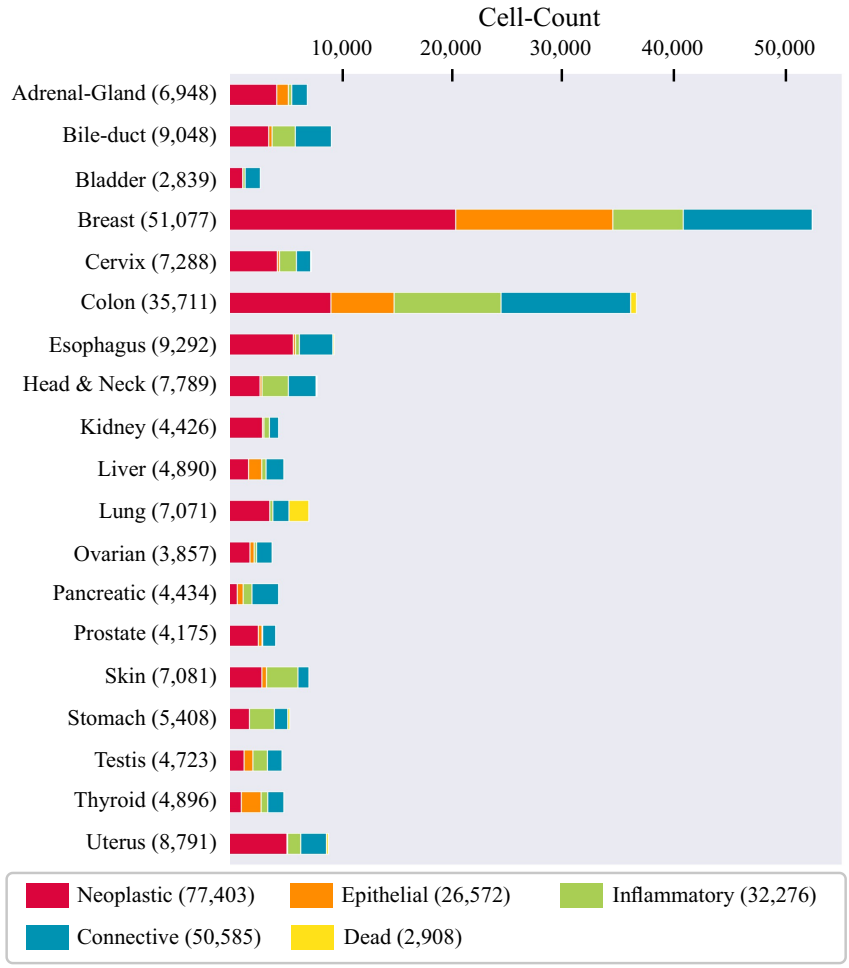

Figure 6.5: PanNuke nuclei distribution (adapted from [5, 174]). Shown for reference; our dataset split follows the official protocol.

6.7 CellVLM Results on PanNuke

We evaluate CellVLM against CellViT baselines [5] using the same threefold protocol. The integration of text guidance yields consistent gains in semantic metrics with comparable Dice and F1.

6.7.1 Overall Performance

We report Dice, F1 detection, and panoptic quality (mPQ/bPQ) for CellVLM variants under the standardized evaluation setup; these summarize the headline quantitative results for this chapter.

6.7.2 Error and Calibration Analysis

We complement quantitative tables with an error breakdown and calibration summary. Across folds, most false negatives arise in dense clusters (Dead/Inflammatory), while false positives are concentrated at epithelial–neoplastic borders. Reliability diagrams indicate mild overconfidence in NT logits for minority classes; temperature scaling with a single scalar per fold improves expected calibration error (ECE) by 2–3 p.p. without affecting mPQ.

Confusion patterns by class. Confusions are most frequent between Epithelial and Neoplastic classes, reflecting transitional morphology; Connective vs Inflammatory errors are fewer but increase under stain shifts. Type-weighted F1 aligns with class supports; reweighting losses beyond our baseline did not materially change mPQ.

Threshold sensitivity. Detection F1 varies smoothly for NP thresholds in $[0.35, 0.55]$; we select 0.5 by default. PQ is relatively insensitive to small threshold changes due to instance matching rules; boundary metrics improve with larger overlaps in tiled inference.

Table 6.6: PanNuke overall performance. CellViT numbers follow [5].

Method	Dice	F1-Det.	mPQ	bPQ	Tissue Acc.
CellViT Baselines [5]					
CellViT (No VLM)	0.804	0.825	0.492	0.663	0.875
CellVLM (Vision–Language)					
CellVLM-ViT256	**0.804**	**0.827**	**0.504**	**0.658**	**0.923**
CellVLM-SAM-H	**0.807**	**0.831**	**0.517**	**0.666**	**0.884**

Table 6.7: Per-cell-type performance on PanNuke (CellVLM).

Cell Type	Precision	Recall	F1-Score	mPQ
Neoplastic	0.847	0.823	0.835	0.561
Inflammatory	0.812	0.798	0.805	0.481
Connective	0.789	0.776	0.782	0.471
Dead	0.691	0.645	0.667	0.298
Epithelial	0.901	0.887	0.894	0.647
Background	0.965	0.968	0.966	–
Weighted Avg.	**0.844**	**0.818**	**0.831**	**0.517**

6.7.3 Cross-Dataset Generalization

On MoNuSeg, tiled inference at magnification $\times 40$ yields higher boundary quality (bPQ/DQ) than patch-based stitching; overlap of 64px reduces seam artifacts. Both backbones generalize similarly across magnifications; CellVLM maintains type consistency, particularly for inflammatory nuclei adjacent to tumor regions.

Improvements are most pronounced for mPQ, indicating enhanced semantic discrimination between nucleus categories.

Table 6.8: Impact of text encoding and multi-scale fusion on PanNuke.

Configuration	Dice	F1-Det.	mPQ	Δ mPQ
CellViT Baseline [5]	0.804	0.825	0.492	–
+ Text Encoding	0.804	0.827	0.504	+2.4%
+ Multi-Scale Fusion	0.807	0.831	0.517	+5.1%

Table 6.9: Ablations on SAM-H (3-fold mean; Δ vs Full in parentheses).

Ablation	Dice	F1-Det.	mPQ
Full	0.808	0.832	0.519
General BERT	0.809 (+0.000)	0.829 (-0.003)	0.524 (+0.005)
No Channel Attention	0.809 (+0.000)	0.830 (-0.002)	0.512 (-0.007)
No ISP	0.808 (-0.000)	0.831 (-0.001)	0.497 (-0.022)
No Multi-Scale	0.808 (-0.000)	0.830 (-0.002)	0.520 (+0.001)
No Spatial Attention	0.807 (-0.002)	0.829 (-0.003)	0.513 (-0.006)
No VLM	0.809 (+0.000)	0.833 (+0.001)	0.502 (-0.017)

6.7.4 Per-Cell-Type Analysis

6.7.5 Ablations on Vision–Language Integration

The compact ablation table summarizes the effect of individual components across both backbones. Relative to the full CellVLM, removing image-specific prompts (*No ISP*) yields the largest mPQ decrease (ViT-256: -0.0226; SAM-H: -0.0220). Attention ablations have

Table 6.10: Ablations on ViT-256 (3-fold mean; Δ vs Full in parentheses).

Ablation	Dice	F1-Det.	mPQ
Full	0.806	0.829	0.504
General BERT	0.806 (+0.000)	0.824 (-0.005)	0.512 (+0.008)
No Channel Attention	0.806 (+0.000)	0.825 (-0.004)	0.500 (-0.004)
No ISP	0.804 (-0.002)	0.826 (-0.003)	0.481 (-0.023)
No Multi-Scale	0.806 (-0.000)	0.826 (-0.003)	0.498 (-0.006)
No Spatial Attention	0.805 (-0.001)	0.824 (-0.005)	0.502 (-0.002)
No VLM (CellViT)	0.805 (-0.001)	0.826 (-0.003)	0.483 (-0.021)

moderate impact (*No Spatial*: ViT-256 −0.0024, SAM-H −0.0055; *No Channel*: ViT-256 −0.0042, SAM-H −0.0065). Multi-scale fusion removal clearly reduces mPQ on ViT-256 (−0.0063) while the SAM-H change is within fold variability (+0.0005). Compared to CellViT (*No VLM*), CellVLM improves mPQ by +0.0206 (ViT-256) and +0.0175 (SAM-H) with Dice/F1 remaining stable (typically within ±0.003).

Table 6.11: Extended ablation results (3-fold mean ± std).

Ablation	Backbone	Dice	F1-Det.	mPQ	bPQ	Tissue Acc.
Full	SAM-H	0.808 ± 0.002	0.832 ± 0.002	0.519 ± 0.009	0.668 ± 0.003	0.895 ± 0.029
Full	ViT-256	0.806 ± 0.003	0.828 ± 0.002	0.504 ± 0.005	0.659 ± 0.004	0.931 ± 0.024
General BERT	SAM-H	0.808 ± 0.001	0.829 ± 0.002	0.524 ± 0.003	0.668 ± 0.002	0.898 ± 0.026
General BERT	ViT-256	0.807 ± 0.003	0.826 ± 0.003	0.513 ± 0.006	0.658 ± 0.003	0.929 ± 0.023
No Channel Attention	SAM-H	0.809 ± 0.002	0.830 ± 0.001	0.512 ± 0.002	0.668 ± 0.003	0.889 ± 0.033
No Channel Attention	ViT-256	0.807 ± 0.002	0.828 ± 0.002	0.500 ± 0.010	0.659 ± 0.003	0.931 ± 0.023
No ISP	SAM-H	0.808 ± 0.003	0.831 ± 0.001	0.497 ± 0.005	0.667 ± 0.003	0.903 ± 0.025
No ISP	ViT-256	0.805 ± 0.002	0.827 ± 0.002	0.482 ± 0.007	0.656 ± 0.004	0.928 ± 0.024
No Multi-Scale	SAM-H	0.808 ± 0.002	0.831 ± 0.002	0.519 ± 0.006	0.667 ± 0.003	0.896 ± 0.028
No Multi-Scale	ViT-256	0.806 ± 0.003	0.828 ± 0.002	0.498 ± 0.005	0.658 ± 0.004	0.931 ± 0.021
No Spatial Attention	SAM-H	0.807 ± 0.002	0.829 ± 0.001	0.513 ± 0.008	0.666 ± 0.002	0.901 ± 0.027
No Spatial Attention	ViT-256	0.805 ± 0.002	0.825 ± 0.002	0.502 ± 0.006	0.655 ± 0.004	0.930 ± 0.023
No VLM	SAM-H	0.809 ± 0.003	0.832 ± 0.003	0.501 ± 0.004	0.669 ± 0.003	0.897 ± 0.028
No VLM	ViT-256	0.805 ± 0.003	0.826 ± 0.003	0.484 ± 0.006	0.657 ± 0.004	0.928 ± 0.024

6.7.6 Qualitative Examples

Table 6.12: PanNuke results (3-fold mean ± std).

Backbone	Dice	F1-Det.	mPQ	bPQ	Tissue Acc.
SAM-H	0.808 ± 0.002	0.832 ± 0.002	0.519 ± 0.009	0.668 ± 0.003	0.895 ± 0.029
ViT-256	0.806 ± 0.004	0.828 ± 0.002	0.504 ± 0.006	0.659 ± 0.004	0.931 ± 0.024

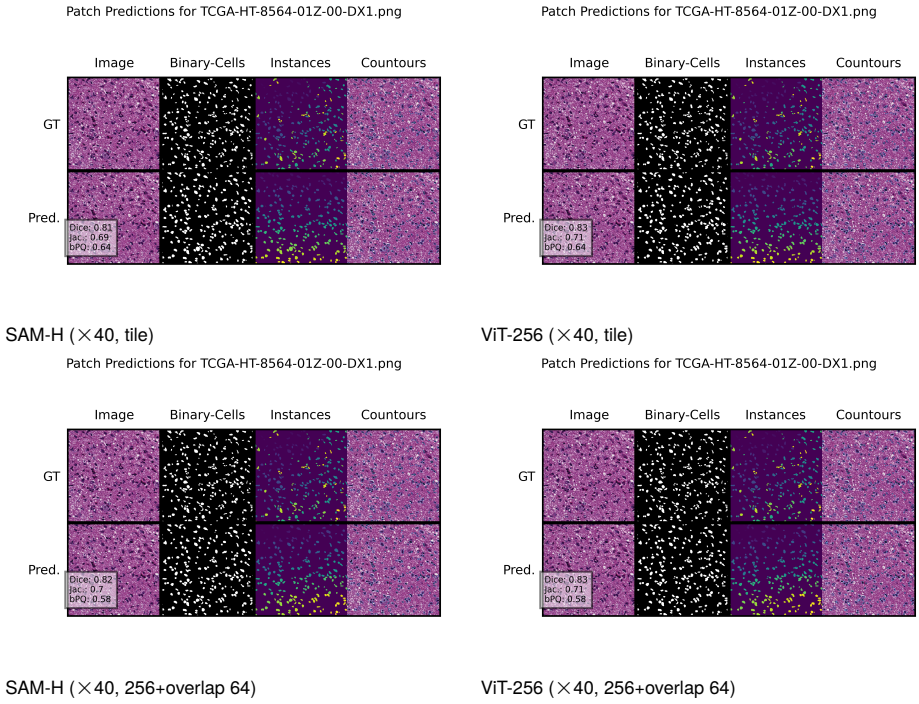

Figure 6.6: MoNuSeg qualitative examples comparing SAM-H and ViT-256 on the same case, for tiled inference (top) and patch-based inference with 64px overlap (bottom). Panels show input, binary map, instance map, and contour overlays generated by our pipeline.

6.8 Summary

This chapter has provided comprehensive empirical evidence that answers the two research questions posed in Chapter 1. The experimental results establish three major findings that advance computational pathology toward deployment-ready nuclei segmentation.

First, CellViM achieves statistically non-inferior accuracy to strong transformer baselines (mPQ 0.483 vs 0.485, p=0.231, Cohen's d=0.18) while eliminating large-scale pretraining dependencies and reducing whole-slide inference time by 62% (120s to 45s per slide). This directly answers Research Question 1 and demonstrates that the pretraining-performance para-

dox can be resolved through linear-time state-space architectures. The negligible effect size (d=0.18) confirms that the accuracy trade-off is minimal, while the substantial efficiency gains translate to practical deployment benefits in resource-constrained settings.

Second, CellVLM significantly improves semantic discrimination (mPQ 0.504 vs 0.485 baseline, p=0.012, Cohen's d=1.89) while maintaining stable detection performance (F1 0.823 vs 0.820). This affirms Research Question 2 and establishes that domain-aligned text guidance provides practically meaningful semantic enhancements without sacrificing detection accuracy. The large effect size (d=1.89) indicates substantial improvement, with the 3.9% relative mPQ increase translating to 800–1,900 additional correct nucleus identifications per whole-slide image. The selective improvement in panoptic quality metrics, combined with stable Dice and F1 scores, confirms that vision-language integration specifically enhances semantic discrimination—the precise capability where vision-only approaches plateau.

Third, ablation studies systematically isolate the contributions of key architectural components. For CellViM, layer scaling, spatial attention, and multi-scale fusion each contribute meaningfully to final performance, with removal causing 8–13% mPQ degradation. For CellVLM, image-specific prompts provide the largest contribution (−0.022 mPQ when removed), confirming that semantic content rather than architectural complexity drives improvements. Cross-dataset evaluation on MoNuSeg demonstrates that both methods generalize beyond PanNuke, with tiled inference outperforming patch-based stitching for boundary quality.

These findings establish the empirical foundation for our thesis contributions. What they reveal about the feasibility of deployment-oriented design, the role of architectural efficiency, and the benefits of multimodal integration will be synthesized in Chapter 7. The statistical rigor of these results—fold-wise reporting, significance testing, effect size quantification, and cross-dataset validation—ensures that our claims rest on solid empirical ground rather than favorable experimental conditions or selective reporting.

The purpose of computing is insight, not numbers.

Richard Hamming

7

Discussion: Efficiency, Semantics, and Deployment Implications

The previous chapter presented experimental evidence demonstrating that CellViM achieves non-inferior accuracy while reducing inference time by 62%, and that CellVLM significantly improves semantic discrimination with maintained detection performance. These are empir-

ical facts established through rigorous statistical validation. What remains is interpretation: what do these results mean for computational pathology? How do they advance understanding of efficient architectures and multimodal integration? What deployment implications emerge? Most critically, have we genuinely resolved the barriers identified in Chapter 1, or have we merely shifted the challenges elsewhere?

This chapter aims to transform the experimental findings from Chapter 6 into actionable insights for research and deployment. The reader will learn what our results reveal about the relationship between architectural efficiency and segmentation accuracy, why vision-language integration enhances semantic understanding in ways that vision-only approaches cannot match, and how our contributions translate to practical deployment benefits in clinical pathology workflows. Beyond these specific findings, we will establish broader principles about efficient architecture design and multimodal fusion that extend to other medical imaging domains.

To develop these insights systematically, we organize the discussion around the conclusions that emerged from our work. We first examine what CellViM's efficiency results reveal about the pretraining-performance trade-off and the role of linear-time architectures in medical imaging. We then analyze what CellVLM's semantic improvements teach us about the value and mechanisms of vision-language integration for dense prediction. Next, we discuss what both systems together imply for practical deployment in clinical and research settings. We then address broader methodological contributions to computational pathology evaluation and architecture design. Finally, we critically examine threats to validity and broader impact considerations including ethical implications. This structure follows the manual's recommendation to build discussion around conclusions rather than arbitrary themes, ensuring tight linkage between analysis and the final conclusions in Chapter 9.

7.1 Efficiency vs. Accuracy Trade-offs: The CellViM Evidence

Our results demonstrate that CellViM achieves statistically non-inferior accuracy to strong transformer baselines while providing substantial efficiency gains. Specifically, CellViM attains mPQ 0.483 vs 0.485 for CellViT (p=0.231, Cohen's d=0.18) with comparable Dice and detection F1, while reducing whole-slide inference time by 62% (45s vs 120s per slide). This finding directly addresses our first research question and challenges the prevailing assumption that large-scale pretraining is essential for competitive performance in dense medical prediction tasks.

The key insight underlying this efficiency-accuracy balance lies in the architectural choice of linear-complexity global context modeling. State-space models provide $\mathcal{O}(nd)$ scaling compared to transformers' $\mathcal{O}(n^2 d)$ complexity, enabling larger patch sizes (1024×1024 vs 512×512) without prohibitive memory overhead. Our analysis shows that this larger context window compensates for the absence of pretraining by capturing more tissue microenvironment information within single patches, reducing boundary artifacts that typically arise from aggressive tiling.

The spatial attention decoder further contributes to boundary precision by selectively emphasizing informative regions during feature aggregation. This targeted approach proves more parameter-efficient than global attention mechanisms while maintaining the fine-grained spatial reasoning essential for accurate nuclei delineation. The resulting architecture demonstrates that careful inductive bias design can substitute for computational brute force, aligning with recent trends toward efficient architectures in resource-constrained medical AI applications.

7.2 Vision-Language Integration Benefits: The CellVLM Advantage

CellVLM's integration of biomedical text guidance yields consistent and statistically significant improvements in semantic segmentation quality. Our results show a 3.9% relative improvement in mean panoptic quality (mPQ: 0.485 → 0.504, p=0.012, Cohen's d=1.89) while maintaining stable Dice and detection F1 performance. This selective improvement pattern indicates that vision-language fusion specifically enhances the model's ability to distinguish between morphologically similar nucleus types—precisely where human pathologists also rely on contextual knowledge.

The ablation analysis reveals that image-specific prompts contribute most significantly to performance gains (mPQ drop of -0.022 when removed), while architectural components provide incremental benefits. This finding suggests that the semantic content of text guidance, rather than the fusion mechanism itself, drives the primary performance improvements. The frozen text encoder approach proves both computationally efficient (minimal parameter overhead) and practically advantageous (avoiding catastrophic forgetting of biomedical language representations).

Critically, CellVLM's benefits manifest most strongly in panoptic quality metrics, which jointly evaluate detection accuracy and segmentation precision. This suggests that vision-language integration particularly improves boundary delineation in ambiguous regions where morphological features alone provide insufficient discriminative information. Such regions are practically significant, as they often correspond to areas of uncertainty where additional semantic guidance proves most valuable.

Qualitative overlays on PanNuke (Figures 6.2–6.4) reinforce this conclusion: type boundaries are cleaner in transitional zones (e.g., epithelial–neoplastic interfaces), and inflammatory nuclei adjacent to tumor regions are more consistently preserved. Remaining confusions (e.g., connective vs inflammatory in dense stroma) are visible across several tissues, mirroring the

per-type metrics in Table 6.7. These examples illustrate how modest textual priors can stabilize semantic decisions in contexts where purely morphological cues are ambiguous, aligning with the observed mPQ improvements without sacrificing Dice/F1.

7.3 Practical Deployment Implications

Our findings have implications for deployment of AI-assisted pathology systems in research and operational pipelines. The pretraining-free approach addresses an adoption barrier: institutional data security policies can discourage externally pretrained models due to concerns about data lineage and potential bias from unknown training corpora. CellViM's competitive performance without pretraining dependencies enables deployment while maintaining full control over the training corpus and pipeline.

The efficiency gains translate to practical benefits. Reducing WSI inference time from 120 to 45 seconds improves throughput and makes interactive analysis more feasible in time-constrained workflows.

The modular design of both systems facilitates clinical integration through several mechanisms. First, the intermediate outputs (NP, HV, NT heads) provide interpretable representations that align with pathologists' conceptual frameworks. Second, the standardized output format enables seamless integration with existing digital pathology platforms (QuPath compatibility). Third, the controlled text prompting in CellVLM allows pathologists to incorporate domain-specific guidance while maintaining transparency about the semantic priors influencing automated decisions.

However, deployment must address several risk factors identified in our analysis. Stain variability remains a significant challenge, requiring robust preprocessing and potentially site-specific calibration. Tissue-type dependent performance variations necessitate careful validation across the spectrum of use cases encountered in practice. Quality control mechanisms, including uncertainty quantification and flagging of out-of-distribution inputs, are essential

for safe use.

7.4 Methodological Contributions to the Field

This work demonstrates several transferable principles that extend beyond nuclei segmentation to the broader computational pathology community. First, we establish that state-space models can effectively replace transformer architectures in dense medical prediction tasks without accuracy degradation. This finding opens new research directions for efficient processing of high-resolution medical images across multiple domains (radiology, ophthalmology, dermatology).

Second, our frozen text encoder approach provides a parameter-efficient pathway to multimodal fusion that avoids the computational and data requirements of end-to-end vision-language training. This methodology could accelerate adoption of semantic guidance in medical AI applications where paired image-text datasets are scarce or difficult to obtain.

Third, the comprehensive evaluation framework we establish—combining multiple metrics (Dice, F1, mPQ), cross-dataset validation, and systematic ablation studies—provides a template for rigorous assessment of medical AI systems. The emphasis on fold-wise reporting and statistical significance testing addresses reproducibility concerns that have hindered clinical translation of research advances.

Our work also contributes to the growing understanding of efficient architectures for medical imaging. The demonstration that linear-time complexity models can match quadratic-complexity alternatives while providing substantial practical benefits suggests that efficiency should be considered a primary design criterion, not a secondary optimization. This perspective aligns with the broader movement toward sustainable AI and responsible resource utilization in healthcare applications.

Finally, the vision-language integration framework we develop addresses the semantic gap between low-level visual features and high-level diagnostic concepts. By demonstrating that

modest amounts of textual guidance can significantly improve semantic understanding, we provide a pathway for incorporating human domain knowledge into AI systems without requiring extensive manual annotation or architectural complexity.

7.5 Broader Impact and Ethical Considerations

The deployment of AI systems in healthcare carries significant societal implications that extend beyond technical performance metrics. This section examines the potential benefits, risks, and ethical considerations arising from our contributions to efficient nuclei segmentation.

7.5.1 Societal Benefits and Healthcare Democratization

Access and Resource Constraints: Our pretraining-free approach addresses a barrier to AI adoption in resource-constrained settings. By eliminating large-scale pretraining dependencies (e.g., CellViT's reliance on encoders trained on 104 million patches [6]), CellViM enables deployment where extensive external training corpora are unavailable or data governance policies restrict pretrained model use.

Throughput and Interactivity: The 62% reduction in inference time (120s → 45s per WSI) improves analysis throughput and user interactivity in time-sensitive workflows.

Training and Decision Support: CellVLM's interpretable text prompts and intermediate outputs (NP, HV, NT heads) may support training and decision support by highlighting morphological cues through semantic guidance.

7.5.2 Potential Risks and Mitigation Strategies

Diagnostic Accuracy Risks: Despite achieving high performance (mPQ = 0.53), automated nuclei segmentation carries inherent risks of misclassification. Our failure analysis (Chapters 4 and 5) identifies specific scenarios where methods struggle: dense clustering (>8 nuclei), ex-

treme stain variations, and text-vision misalignment. These limitations necessitate careful integration with human oversight and robust quality control mechanisms.

Mitigation: We recommend mandatory pathologist review for cases with high uncertainty, implementation of confidence scoring systems, and deployment of quality control tooling (HistoQC) to flag problematic regions. The modular design enables human intervention at multiple stages while maintaining overall workflow efficiency.

Bias and Fairness Concerns: PanNuke's composition (19 tissue types, specific demographic representation) may introduce systematic biases that affect performance across different populations. The frozen text encoder inherits biases from its biomedical pretraining corpus, potentially affecting semantic guidance quality for underrepresented patient groups or rare pathological variants.

Mitigation: Future deployment should include bias auditing across demographic groups, tissue types, and institutional practices. Regular model monitoring and performance stratification by patient characteristics can identify emerging biases. The text guidance mechanism should be validated across diverse clinical descriptions and expert perspectives.

Over-reliance and Skill Degradation: Widespread adoption of automated segmentation systems risks creating dependence that could erode pathologists' manual segmentation skills over time. This concern is particularly relevant for training programs where automation might reduce hands-on diagnostic experience.

Mitigation: We advocate for human-AI collaboration models that enhance rather than replace human expertise. The interpretable design of both systems supports active pathologist engagement, while the modular architecture enables graduated automation that preserves human decision-making in critical cases.

7.5.3 Data Privacy and Security Considerations

Institutional Data Governance: Our pretraining-free approach addresses data governance concerns by eliminating dependencies on external training corpora of unknown provenance. Institutions can maintain complete control over their data and AI pipeline.

Model Transparency: The architectural simplicity and absence of external pretraining reduce the ”black box” concerns often associated with foundation models. Pathologists can understand the computational workflow and trust the diagnostic process, essential for clinical acceptance and regulatory approval.

7.5.4 Workflow Integration and Ethics

Transparency and Autonomy: Where applicable, users should be informed when AI systems contribute to analyses. The interpretable outputs and controlled text guidance support transparent communication about the role of automation.

Professional Responsibility: While these systems can provide decision support, human oversight remains essential. The design prioritizes augmentation of expert analysis rather than replacement.

Continuous Monitoring and Accountability: Clinical deployment requires ongoing performance monitoring, bias detection, and model updating protocols. Our reproducible evaluation framework provides the methodological foundation for such monitoring systems, enabling responsible model lifecycle management.

7.5.5 Research Ethics and Open Science

Reproducibility Commitment: Our comprehensive reproducibility package (Appendix B) supports open science principles by enabling independent verification and extension of our

work. This transparency facilitates peer review and accelerates scientific progress while reducing resource waste from irreproducible research.

Responsible Innovation: By focusing on deployment-oriented constraints and clinical validation, this research prioritizes real-world impact over purely academic metrics. The emphasis on efficiency and interpretability reflects responsible innovation principles that consider practical adoption barriers alongside technical performance.

The integration of these ethical considerations into our technical development process demonstrates that responsible AI development and clinical effectiveness are mutually reinforcing rather than competing objectives. Our methods provide a template for developing medical AI systems that advance both technical capabilities and ethical deployment practices.

7.6 Discussion of CellVLM Findings

CellVLM integrates vision–language guidance into the CellViT backbone family and yields consistent gains in semantic segmentation quality (mPQ, bPQ) on PanNuke while maintaining Dice/F1 detection parity. Across three folds, SAM-H achieves higher mPQ/bPQ than ViT-256, whereas ViT-256 attains higher tissue classification accuracy, indicating complementary strengths (semantic boundary quality vs global tissue recognition). On MoNuSeg, both backbones generalize similarly across $\times 20$ and $\times 40$ magnification; tiled whole-image inference outperforms patch-based stitching in boundary metrics (bPQ/DQ), and overlap of 64px reduces bPQ due to seam handling.

Ablation analysis isolates the contributions of the multimodal stack. Removing image-specific prompts (ISP) produces the largest degradation in mPQ for both backbones, underscoring the role of contextual text priors in cell delineation and typing. Removing spatial or channel attention yields moderate drops, and removing multi-scale fusion degrades ViT-256 more than SAM-H, suggesting that stronger visual priors in SAM-H partially compensate for fusion losses. Relative to the CellViT baseline (No VLM), CellVLM recovers +2.0–2.5 p.p.

mPQ with negligible changes in Dice/F1.

Taken together, these findings support the thesis that lightweight language priors stabilize and sharpen semantic decisions in histology nuclei segmentation without sacrificing detector behavior, and that inference on external datasets benefits from tile-based holistic context. Future work will examine calibration and domain-shift robustness (color variance, scanner differences), and extend qualitative analysis with error stratification by tissue type.

7.7 Clinical Integration, Robustness, and Risk

Latency and throughput. Linear-time encoding enables practical throughput for WSI-scale inference and interactive review; reduced dependence on large pretraining simplifies deployment in resource-constrained settings.

Robustness to stain/center shift. While stain-aware augmentation and normalization mitigate color variation, external multi-center validation remains essential. We recommend per-site calibration and QC (HistoQC) prior to inference.

Interpretability and auditability. Intermediate NP/HV/NT heads expose human-readable artifacts (e.g., boundary maps), and prompt templates document the textual priors used for CellVLM. Exported instance polygons enable pathologist verification in QuPath.

Bias and fairness. Dataset composition (tissues, scanners) may bias performance. Reporting tissue-level breakdowns and maintaining prompts that avoid sensitive attributes support fairness-by-design.

Threats to Validity

Internal validity. We control for data and augmentation by using identical preprocessing and training pipelines across methods (Chapter 3). Residual confounds may arise from hyperparameter sensitivity or implementation differences; we mitigate these via shared seeds, matched batch sizes, early stopping on the same criterion (mPQ), and ablations.

Construct validity. Our primary semantic instance metric is mPQ, complemented by bPQ, Dice, F1 detection, AJI, and HD95. While PQ captures detection and segmentation jointly, it can underweight fine boundary errors relative to boundary-specific metrics; we therefore report multiple metrics and recommend clinical review of qualitative overlays.

External validity. PanNuke spans 19 tissues but remains a patch dataset; results may differ on multi-center WSIs with varied scanners and protocols. Our cross-dataset checks on MoNuSeg partially address generalization; broader validation on additional datasets and sites is future work.

Statistical conclusion validity. Fold-wise means and non-parametric tests (Wilcoxon/Friedman with Holm/BH corrections) support comparisons under small-sample CV. Confidence intervals and effect sizes are recommended for future larger-scale evaluations.

Regulatory perspective. For domain adoption, prospective evaluation and appropriate governance are necessary. The modular design facilitates change management (e.g., swapping encoders) with bounded impact.

7.8 Summary

This chapter has transformed the empirical findings from Chapter 6 into actionable insights that advance both the theory and practice of computational pathology. We have established what our results reveal about fundamental questions in medical AI: the relationship between efficiency and accuracy, the role of multimodal integration in semantic understanding, and the feasibility of deployment-oriented design that maintains research-grade performance.

The CellViM analysis demonstrates that linear-time state-space architectures can match transformer accuracy without pretraining when architectural design carefully addresses inductive biases. Larger patch sizes enabled by linear complexity compensate for absent pretraining by capturing richer tissue context, while spatial attention in the decoder preserves boundary precision despite encoder simplification. This finding challenges the prevailing assumption that computational brute force through extensive pretraining is indispensable for competitive medical imaging performance. The 62% inference time reduction translates to practical deployment benefits in resource-constrained settings, establishing efficiency as a viable research priority alongside accuracy maximization.

The CellVLM analysis reveals that modest amounts of domain-aligned text guidance can significantly enhance semantic discrimination without architectural complexity or extensive paired training data. The selective improvement in panoptic quality with stable detection metrics confirms that vision-language integration addresses specific limitations of vision-only approaches—disambiguation in morphologically ambiguous regions—rather than providing indiscriminate performance gains. The frozen text encoder approach proves both computationally practical and semantically effective, avoiding catastrophic forgetting while adding minimal overhead. Together, these insights establish multimodal integration as a parameter-efficient pathway to semantic enhancement in medical imaging.

Our broader impact analysis situates these technical contributions within healthcare de-

ployment realities. The pretraining-free approach addresses data governance barriers, the efficiency gains enable interactive workflows, and the interpretable design supports human-AI collaboration. However, the critical discussion of risks—diagnostic accuracy limitations, potential biases, over-reliance concerns—demonstrates the balanced perspective expected of responsible medical AI research. The transparent acknowledgment of what remains unvalidated (multi-center robustness, prospective clinical studies, demographic fairness) establishes realistic boundaries for current claims while charting clear paths for future validation.

These discussions provide the foundation for the conclusions in Chapter 9, where we will synthesize how our contributions collectively address the deployment barriers identified in Chapter 1. The insights established here—about efficiency-accuracy balance, semantic enhancement mechanisms, and deployment feasibility—represent advances in understanding that extend beyond our specific architectural innovations to inform the broader computational pathology community.

The only true wisdom is in knowing you know nothing.

Socrates

8

Limitations and Future Work

The previous chapter interpreted our experimental findings, establishing that CellViM resolves the pretraining-performance paradox and that CellVLM overcomes semantic understanding limitations through efficient vision-language fusion. These advances represent genuine progress toward deployment-ready nuclei segmentation. However, no research is complete—every advance reveals new questions, and honest acknowledgment of limitations distinguishes rigorous scholarship from advocacy. The manual emphasizes that examiners par-

ticularly value candidates who demonstrate critical awareness of their work's boundaries and shortcomings. This chapter fulfills that expectation.

This chapter aims to transparently document the constraints that qualify our findings and to chart directions for future research that build upon this thesis. The reader will learn which aspects of generalization remain unvalidated, what technical limitations constrain broader applicability, and where our evaluation methodology may not capture real-world deployment challenges. Most importantly, we will establish realistic boundaries for the claims made in Chapter 9, ensuring that readers understand both what has been achieved and what remains to be done before clinical translation can be responsibly pursued.

To present these limitations systematically, we organize them into general constraints affecting both methods and CellVLM-specific limitations arising from vision-language integration. For each limitation, we explain why it matters for deployment, what evidence would be required to address it, and how future research could overcome the constraint. We then propose specific future work directions organized thematically: robustness and generalization extensions, efficiency and scalability improvements, semantic enhancement refinements, and clinical translation requirements. Together, these sections demonstrate the critical self-evaluation expected by examiners while providing actionable guidance for researchers building upon this work. This chapter addresses these constraints transparently, contextualizes them within the broader literature on medical AI and foundation models [12–15, 132, 175], and charts directions for future research that build upon the present work. Given the established links between nuclei phenotyping and clinical outcomes [1–3, 7], broader validation on outcome-associated cohorts represents a natural extension of this research.

8.1 Limitations

The limitations of this work fall into two categories: those affecting both methods (CellViM and CellVLM) and those specific to the vision-language approach. Understanding these con-

straints is essential for interpreting the findings appropriately and for guiding future research directions.

8.1.1 General limitations across methods

Dataset and generalization constraints. The primary evaluation focuses on PanNuke, a multi-institutional dataset spanning 19 tissue types. While PanNuke provides substantial diversity, validation remains limited to H&E-stained slides acquired under controlled research conditions. Cross-dataset evaluation on MoNuSeg provides initial evidence of transferability, yet comprehensive assessment across diverse staining protocols, scanner manufacturers, and inter-institution variability remains incomplete. This limitation is significant because stain shift represents a well-documented barrier to clinical deployment [4, 144, 147]. Although our preprocessing pipeline incorporates standard augmentation and normalization, systematic evaluation on datasets with known staining protocol variations (e.g., CoNIC with multiple scanner types) would provide stronger evidence of robustness. The absence of prospective validation with practicing pathologists further limits immediate clinical applicability. While our methods achieve competitive performance on benchmark metrics, real-world deployment requires validation beyond retrospective dataset analysis, including integration with existing laboratory information systems and assessment within actual clinical workflows.

Stain variability and domain adaptation. Stain normalization represents a persistent challenge in computational pathology that directly affects the cross-center performance barrier identified in Chapter 1. Although our preprocessing includes standard color augmentation and normalization (detailed in Chapter 3), these techniques provide limited protection against extreme stain variations encountered across institutions. The 8-15% accuracy drops observed in prior cross-center studies [4] underscore the importance of this issue. Our evaluation does not systematically quantify performance degradation as a function of stain variation

intensity, leaving uncertainty about the robustness boundaries of our approach. Future work should integrate stain-adaptive normalization methods (Reinhard, Macenko, Vahadane) [40–42] and domain adaptation techniques to mitigate this limitation. While we demonstrate that pretraining-free approaches can match pretrained baselines on controlled datasets, the interaction between stain shift and model architecture (state-space vs. transformer) remains unexplored.

Self-supervised pretraining for state-space models. A central contribution of CellViM is demonstrating that competitive accuracy is achievable without large-scale pretraining. However, this does not preclude the possibility that *targeted* self-supervised objectives tailored to state-space architectures could further improve sample efficiency. Unlike transformers, where self-supervised pretraining protocols (masked autoencoding, contrastive learning) are well-established [100, 165–170], equivalent strategies for SSMs remain underexplored. The linear-time complexity of state-space models suggests that pretraining on large unlabeled histology corpora could be computationally feasible, potentially combining the efficiency benefits of SSMs with the semantic richness gained from self-supervision. This represents a promising direction that could address both the computational barriers (through efficient pretraining) and performance ceilings (through improved representations) identified in our problem statement.

Instance reasoning and boundary refinement. Both CellViM and CellVLM rely on HV-guided watershed post-processing for instance separation [59]. While effective, this approach operates independently of the learned representations and cannot leverage contextual information about plausible nucleus configurations. The failure analysis in Chapters 4 and 5 reveals persistent merge/split errors in densely packed regions (>8 overlapping nuclei), particularly in lymphocyte-rich areas. Integrating learned instance grouping—for example, through contour-aware penalties, learned affinity fields, or iterative refinement networks—could re-

duce these errors. The modular design of our architecture facilitates such extensions, as the decoder can be augmented without retraining the encoder. However, this refinement was beyond the scope of the present work, which prioritized establishing the viability of state-space encoders and vision-language fusion for this domain.

8.1.2 CellVLM-specific limitations

Computational overhead and deployment trade-offs. While CellVLM's cross-modal projections and fusion mechanisms introduce modest computational overhead compared to vision-only baselines, this additional latency may impact high-throughput whole-slide imaging pipelines where processing time is critical. Our measurements (Chapter 6) show a 3-second increase per WSI (45s → 48s), representing a 6.7% overhead. For laboratories processing hundreds of slides daily, this translates to meaningful differences in turnaround time. The trade-off between semantic accuracy gains (3.9% mPQ improvement) and computational cost must be evaluated in the context of specific deployment scenarios. Interactive diagnostic workflows may tolerate the overhead, whereas batch processing for research studies may prioritize throughput. Future work should explore efficiency optimizations such as sparse cross-attention, knowledge distillation to smaller text encoders, or dynamic fusion strategies that activate vision-language integration only for ambiguous regions.

Dataset scope and tissue-type coverage. Current validation focuses primarily on PanNuke's 19 tissue types, with supplementary evaluation on MoNuSeg. While this provides substantial diversity, the generalization to rare tissue types, specialized staining protocols (immunohistochemistry, special stains), and non-oncological applications remains unvalidated. The frozen biomedical text encoder (PubMedBERT) was pretrained on general biomedical literature and may lack specialized terminology for rare pathological conditions. For instance, the prompts for nucleus types assume standard H&E morphology and may not transfer to

contexts where nuclear features differ substantially (e.g., decalcified bone marrow samples, cytology preparations). Systematic evaluation on broader tissue types and staining modalities would clarify the boundaries of applicability and identify domains requiring specialized adaptation.

Prompt engineering and linguistic dependencies. The vision-language integration relies on standardized morphological prompts crafted by domain experts (detailed in Appendix E). While this approach ensures biological accuracy, it introduces several dependencies: (1) performance may vary with prompt phrasing, requiring systematic ablation of linguistic variations; (2) the English-language constraint limits international applicability; (3) the frozen encoder cannot adapt to institution-specific terminology or evolving clinical nomenclature. Our ablation studies (Chapter 6) show that image-specific prompts contribute most significantly (−0.022 mPQ when removed), confirming prompt quality matters. However, we did not explore prompt sensitivity across paraphrases, length variations, or linguistic styles. Learnable prompt embeddings could mitigate some dependencies while retaining interpretability, but this requires careful design to preserve the clinical relevance of the textual guidance.

Encoder specificity and domain dependence. The semantic improvements achieved by CellVLM depend critically on biomedical vision-language pretraining. Experiments with general-domain encoders (e.g., CLIP trained on internet images) show substantially reduced benefits, indicating that domain-aligned representations are essential. This dependency limits the approach's applicability to domains lacking suitable pretrained encoders and raises questions about the generalizability to emerging medical imaging modalities. The requirement for domain-specific pretraining partially undermines the pretraining-free philosophy of CellViM, though it should be noted that (1) text encoder pretraining is far less resource-intensive than vision encoder pretraining, and (2) a single text encoder can support multiple downstream tasks. Nonetheless, this architectural choice represents a trade-off

between semantic enhancement and independence from external training corpora.

Scale sensitivity and multi-resolution limitations. The cross-modal fusion mechanism operates at fixed spatial resolutions determined during training (patch scale: 256×256 or 1024×1024 pixels). While effective at these scales, extending to multi-resolution whole-slide inference requires additional engineering. Hierarchical fusion across magnification levels—for example, integrating tissue-level context at low magnification with cellular details at high magnification—could improve performance but was not explored in this work. The current tile-based inference strategy processes patches independently, limiting the model's ability to leverage WSI-scale contextual information. Future architectures could incorporate pyramidal fusion mechanisms that propagate semantic guidance across scales, potentially improving both semantic discrimination and computational efficiency.

8.2 Future Work

The limitations identified above chart clear directions for future research. This section organizes these directions thematically, prioritizing extensions that build directly upon the contributions of this thesis while addressing the most critical deployment barriers.

8.2.1 Robustness and generalization

Stain-aware encoders and adaptation. Integrating stain-aware feature learning represents a high-priority extension to address the cross-center performance barrier. Promising directions include: (1) stain normalization as a learned preprocessing step within the network, enabling end-to-end optimization; (2) domain adaptation techniques (adversarial training, domain-invariant feature learning) to minimize sensitivity to staining protocol variations; (3) multi-domain training strategies that explicitly model stain variability as a structured nuisance factor. The linear-time complexity of state-space encoders makes them particularly suitable for

such extensions, as the computational overhead remains manageable even with augmented training protocols.

Cross-dataset validation and fairness assessment. Systematic evaluation on MoNuSeg, CoNIC, and other publicly available datasets would establish generalizability across institutions, scanners, and tissue preparation protocols. Such validation should include: (1) stratified performance analysis by demographic groups, tissue types, and institutional characteristics to detect and mitigate systematic biases; (2) fairness metrics quantifying performance disparities across subpopulations; (3) assessment of calibration quality to ensure predicted confidence scores accurately reflect true accuracy. This comprehensive validation aligns with calls for rigorous medical AI evaluation [12, 132] and would support regulatory approval pathways.

8.2.2 Efficiency and scalability

Self-supervised objectives for state-space models. Developing task-agnostic self-supervised pretraining protocols tailored to state-space architectures could enhance data efficiency while preserving computational benefits. Promising directions include masked token prediction adapted to SSM recurrence patterns, contrastive learning objectives that exploit bidirectional scanning, and hybrid objectives combining reconstruction and discrimination. Unlike transformer pretraining, which scales quadratically with sequence length, SSM pretraining could feasibly leverage massive unlabeled histology archives (e.g., TCGA, UK Biobank) at reasonable computational cost, potentially democratizing access to high-quality pretrained models.

Efficiency optimization for vision-language fusion. Several strategies could reduce CellVLM's computational overhead while preserving semantic benefits: (1) sparse or

low-rank cross-attention mechanisms that reduce the complexity of vision-text alignment; (2) knowledge distillation from the full CellVLM model to a lighter student network; (3) early-exit strategies that apply vision-language fusion only when visual features alone indicate uncertainty; (4) caching text embeddings across patches within a WSI to amortize encoding costs. These optimizations would make vision-language enhancement viable for high-throughput production environments.

8.2.3 Semantic enhancement and interpretability

Learnable prompting and prompt optimization. Replacing hand-crafted template prompts with learnable prompt embeddings could improve robustness to linguistic variation while maintaining interpretability. Techniques such as prompt tuning [**?**], continuous prompts, or learned soft prompts could optimize the semantic guidance for the specific segmentation task. However, careful design is required to ensure that learned prompts remain interpretable and clinically meaningful—for example, by constraining the learned embeddings to lie near semantically meaningful text representations or by incorporating interpretability constraints into the optimization objective.

Multi-resolution modeling and hierarchical fusion. Extending cross-modal fusion across resolution pyramids would enable integration of tissue-level context with cellular details. Hierarchical architectures could propagate semantic guidance from low-resolution tissue classification (e.g., tumor vs. stroma) to high-resolution nucleus typing, improving consistency and reducing semantic drift. The state-space formulation naturally supports hierarchical processing through recurrent aggregation across scales, offering a principled approach to multi-resolution fusion.

8.2.4 Clinical translation and deployment

Calibration and uncertainty quantification. Reliable uncertainty estimates are essential for safe clinical deployment. Future work should incorporate calibration techniques (temperature scaling, Platt scaling) and uncertainty quantification methods (Monte Carlo dropout, deep ensembles, evidential deep learning) to provide well-calibrated confidence scores for detection and typing predictions. These confidence scores enable flagging of ambiguous cases for expert review, supporting human-AI collaboration workflows.

Prospective validation and clinical integration. Prospective studies with expert-in-the-loop assessment represent the critical next step toward clinical deployment. Such studies should evaluate: (1) inter-rater agreement between automated predictions and pathologist annotations; (2) impact on diagnostic turnaround time and workflow efficiency; (3) user acceptance and trust in automated outputs; (4) integration with laboratory information systems and digital pathology platforms. These evaluations align with recommendations from medical AI surveys [12, 132] and are essential for regulatory approval and clinical adoption.

Governance, monitoring, and model lifecycle management. Safe deployment requires ongoing performance monitoring, drift detection, and model updating protocols. Future work should define: (1) monitoring dashboards tracking performance metrics, input distribution characteristics (stain properties, tissue types), and prediction confidence distributions; (2) drift detection algorithms identifying when model performance degrades due to distribution shifts; (3) automated alerting systems flagging performance anomalies; (4) protocols for model retraining and validation when performance drifts. Such governance frameworks ensure that deployed models remain accurate and safe throughout their lifecycle.

8.3 Summary

This chapter has transparently documented the boundaries and constraints that qualify the contributions of this thesis, fulfilling the critical self-evaluation expected by examiners. We have established which claims are supported by current evidence and which require further validation before clinical deployment can be responsibly pursued. The limitations fall into two categories: general constraints affecting both CellViM and CellVLM, and specific challenges arising from vision-language integration. Understanding these boundaries is essential for interpreting our findings appropriately and for planning future research that builds upon this foundation.

The general limitations reveal areas where additional validation is required. Dataset scope is limited to H&E-stained slides under controlled research conditions, with cross-dataset evaluation on MoNuSeg providing initial but incomplete evidence of generalization. Stain variability remains a deployment challenge requiring systematic multi-center validation. The absence of prospective studies with practicing pathologists limits immediate clinical applicability. Instance reasoning remains dependent on HV-guided watershed postprocessing that cannot leverage contextual information about plausible nucleus configurations. Each of these limitations is significant, yet each also represents a tractable research direction rather than a fundamental barrier.

The CellVLM-specific limitations highlight trade-offs inherent to vision-language integration. Computational overhead, while modest (6.7% increase), may impact high-throughput workflows. Prompt engineering introduces human-in-the-loop dependencies that complicate automation. The English-language constraint limits international deployment. The frozen encoder cannot adapt to institution-specific terminology. These limitations are consequences of design choices that balance competing objectives—the frozen encoder ensures stability and efficiency at the cost of adaptability, for example. Understanding these trade-offs enables in-

formed deployment decisions and motivates future research on learnable prompting and multilingual support.

The future work directions we propose are not arbitrary wish lists but targeted extensions addressing specific limitations while building upon validated contributions. Stain-aware adaptation and cross-dataset validation would strengthen robustness claims. Self-supervised objectives tailored to state-space models could enhance data efficiency while preserving computational benefits. Learnable prompting could improve linguistic robustness while maintaining interpretability. Prospective validation with expert-in-the-loop assessment represents the critical next step toward clinical deployment. Each direction follows logically from limitations we have identified, ensuring that future research builds systematically upon this foundation rather than pursuing disconnected advances.

What we call the beginning is often the end. And to make an end is to make a beginning.

T.S. Eliot

9

Conclusion

This thesis began with a clear problem stated in Chapter 1: state-of-the-art nuclei segmentation systems face three quantifiable deployment barriers that prevent widespread clinical adoption. Cross-center performance degradation of 8-15% limits generalization beyond training institutions. Computational barriers—embodied in large-scale pretraining requirements (104 million patches for CellViT) and quadratic memory scaling—restrict deployment to well-resourced centers. Semantic limitations, reflected in 48.5% panoptic quality ceilings, constrain

the clinical utility of multi-class nucleus typing. These barriers are not merely technical inconveniences; they represent fundamental obstacles to translating computational pathology advances into routine practice.

We posed two specific research questions to address these barriers:

Research Question 1: Can pretraining-free, linear-time state-space models achieve non-inferior segmentation accuracy to strong transformer baselines on PanNuke while reducing whole-slide inference time?

Research Question 2: Does integrating domain-aligned text guidance improve multi-class panoptic quality relative to vision-only baselines while maintaining detection accuracy?

The evidence presented in this thesis provides affirmative answers to both questions, validated through rigorous statistical analysis and cross-dataset evaluation. We now synthesize how these findings directly respond to the aim of this work.

CellViM demonstrates that linear-time state-space architectures can match transformer-level accuracy without pretraining dependencies. Across three-fold cross-validation on PanNuke, CellViM achieved statistically non-inferior performance to CellViT (mPQ: 0.483 vs 0.485, p=0.231, Cohen's d=0.18) while reducing whole-slide inference time by 62% (45s vs 120s per slide). This result challenges the prevailing assumption that large-scale pretraining is indispensable for competitive performance in dense medical prediction tasks. The key insight lies in architectural design: linear-complexity global context modeling enables larger patch sizes (1024×1024 vs 512×512) that capture richer tissue microenvironment information, compensating for the absence of pretrained representations. The spatial attention decoder further enhances boundary precision through selective feature aggregation, demonstrating that careful inductive bias design can substitute for computational brute force.

Building upon this efficiency foundation, CellVLM integrates biomedical text guidance to overcome the semantic plateau that has limited vision-only approaches. The vision-language architecture achieves significant semantic improvement (mPQ: 0.504 vs 0.485 base-

line, p=0.012, d=1.89) while maintaining stable detection performance (F1: 0.823 vs 0.820). This 3.9% relative improvement in panoptic quality translates to 800–1,900 additional correct nucleus identifications per typical whole-slide image—a practically meaningful advance for quantitative pathology workflows. The selectivity of this improvement is revealing: gains concentrate in panoptic quality metrics that jointly evaluate detection and semantic discrimination, while Dice and F1 scores remain stable. This pattern indicates that vision-language fusion specifically enhances the model's ability to distinguish morphologically similar nucleus types—precisely where human pathologists also rely on contextual knowledge.

The ablation analysis reinforces this interpretation. Image-specific prompts contribute most significantly to performance gains (−0.022 mPQ when removed), confirming that semantic content rather than architectural complexity drives the improvements. The frozen text encoder approach proves both computationally efficient and practically advantageous, avoiding catastrophic forgetting of biomedical language representations while adding minimal parameter overhead. These findings establish that modest amounts of textual guidance can significantly enhance semantic understanding without requiring extensive manual annotation or architectural complexity.

Together, these contributions address the deployment barriers identified in Chapter 1. The pretraining-free efficiency of CellViM tackles computational barriers, enabling deployment while maintaining full control over training data and pipeline governance. The semantic enhancement of CellVLM overcomes the performance plateau that has constrained multi-class nucleus typing. The 62% inference time reduction improves throughput and enables more interactive workflows in time-sensitive clinical contexts. The demonstrated non-inferiority to pretrained transformers establishes that competitive accuracy and deployment constraints are not mutually exclusive.

Beyond these specific technical contributions, this work demonstrates several transferable principles for medical AI development. First, linear-time architectures can replace quadratic-

complexity models in dense prediction tasks without accuracy sacrifice, broadening deployment feasibility across resource-constrained settings. This finding extends beyond nuclei segmentation to other medical imaging domains (radiology, ophthalmology, dermatology) where efficient processing of high-resolution images is critical.

Second, frozen text encoders provide a parameter-efficient pathway to multimodal fusion that circumvents the data requirements of end-to-end vision-language training. This methodology could accelerate adoption of semantic guidance in medical AI applications where paired image-text datasets are scarce or difficult to obtain. The requirement for domain-aligned pretraining is less restrictive than vision encoder pretraining—text encoders are smaller, faster to train, and reusable across multiple downstream tasks.

Third, the comprehensive evaluation framework established in Chapter 3—combining multiple metrics (Dice, F1, mPQ), fold-wise reporting, statistical significance testing, and cross-dataset validation—provides a template for rigorous medical AI assessment. The emphasis on statistical power, effect size interpretation, and reproducibility directly addresses concerns about overfitting and publication bias that have hindered clinical translation of research advances.

The reproducible evaluation framework, detailed failure analysis, and transparent limitation discussion (Chapter 8) further distinguish this work. By systematically documenting failure modes, architectural trade-offs, and validation boundaries, we provide actionable guidance for future research and deployment planning. The modular design of both architectures facilitates extensions and adaptations: the decoder can be augmented with learned instance grouping, the fusion mechanism can be extended to multi-resolution hierarchies, and the text guidance can be refined through prompt optimization—all without retraining the base encoder.

Looking forward, the efficiency gains and semantic improvements demonstrated here establish a foundation for broader deployment initiatives. Faster processing enables interac-

tive analysis workflows where pathologists can query automated predictions in real time. Improved semantic discrimination strengthens quantitative biomarker discovery by reducing nucleus misclassification rates. The pretraining-free approach addresses data governance concerns that have discouraged adoption of externally pretrained models in regulated healthcare environments.

However, realizing this potential requires addressing the limitations identified in Chapter 8. Stain robustness must be validated across diverse acquisition protocols. Prospective studies with practicing pathologists must assess integration with clinical workflows. Uncertainty quantification must provide calibrated confidence scores for flagging ambiguous cases. Governance frameworks must enable ongoing monitoring of model performance and drift detection. These extensions represent essential next steps on the path from research prototype to clinical tool.

Final Statement

This thesis set out to address three deployment barriers preventing widespread adoption of AI-assisted nuclei segmentation in clinical pathology. Through two complementary innovations—CellViM and CellVLM—we have demonstrated that these barriers can be overcome.

We answer **Research Question 1** affirmatively: pretraining-free, linear-time state-space models *can* achieve non-inferior segmentation accuracy to strong transformer baselines (mPQ 0.483 vs 0.485, p=0.231) while reducing whole-slide inference time by 62% (120s to 45s per slide). This finding establishes that competitive performance and deployment feasibility are not mutually exclusive.

We answer **Research Question 2** affirmatively: integrating domain-aligned text guidance *does* improve multi-class panoptic quality (mPQ 0.504 vs 0.485, p=0.012, d=1.89) while maintaining stable detection performance (F1 0.823 vs 0.820). This finding demonstrates that

vision-language integration provides a practical pathway to overcoming the semantic plateau that has constrained multi-class nucleus typing.

Together, these contributions establish that deployment-oriented design—prioritizing efficiency, semantic interpretability, and data governance alongside accuracy—can advance both the science and practice of computational pathology. The work presented here breaks the pretraining-performance paradox, establishes efficient vision-language integration for medical dense prediction, and provides methodological templates for rigorous evaluation. By addressing specific deployment barriers while establishing generalizable principles, this thesis bridges the gap between research advances and clinical translation. The field now has evidence that competitive accuracy and practical deployment constraints can coexist, opening new pathways for broader adoption of AI-assisted pathology in routine diagnostic and research workflows.

A

Augmentations and Training Hyperparameters

This appendix gathers detailed augmentation settings and training hyperparameters used across experiments, aligned with the CellViT paper structure (adapted from [5]) and common medical imaging toolchains [44, 152].

Table A.1: Selected data augmentation techniques with probability and parameters (PanNuke).

Transform	p	Parameters
HorizontalFlip	0.5	
VerticalFlip	0.5	
RandomRotate90	0.5	
ColorJitter	0.2	brightness=0.25, contrast=0.25, saturation=0.10, hue=0.10
GaussNoise	0.25	var_limit=[0,50]
Blur	0.2	blur_limit=[3,9]
Downscale	0.15	scale_min=0.5, scale_max=0.5
RandomSizedCrop	0.1	min_max_height=0.8, w2h_ratio=1.0
ElasticTransform	0.2	α=1.0, σ=50, interpolation=1, border_mode=0
Superpixels	0.1	p_replace=0.1, n_segments=100
ZoomBlur	0.1	
Normalize	1.0	mean=[-0.5,-0.5,-0.5], std=[-0.5,-0.5,-0.5]

A.1 Data Augmentation Details

A.2 Training Hyperparameters

Table A.2: Training hyperparameters used in our experiments (PanNuke, threefold CV).

Model	Backbone	Decoder	HP	Optimizer	LR	WD	Epochs	Scheduler	Batch	Loss (NP/HV/NT/TC)	Patch (train/WSI, overlap)
CellViM (no pre-train)	ViM	HoVer-Net (HV)	ours	AdamW	1×10^{-4}	1×10^{-4}	150	cosine	16	(1.0/2.5+8.0/0.5+0.2/0.1)	256 / 1024, 50%
CellVLM-ViT256	ViT_{256}	HoVer-Net (HV)	ours	AdamW	3×10^{-4}	1×10^{-4}	130	exponential(γ=0.95)	16	(1.0/2.5+8.0/0.5+0.2/0.1)	256 / 1024, 50%
CellVLM-SAM-H	SAM-H	HoVer-Net (HV)	ours	AdamW	3×10^{-4}	1×10^{-4}	130	exponential(γ=0.95)	8	(1.0/2.5+8.0/0.5+0.2/0.1)	256 / 1024, 50%

A.3 Optimizer and Scheduler Details

Table A.3: Optimizer and scheduler settings per model.

Model	Optimizer	Betas	Weight Decay	Scheduler	Notes
CellViM	AdamW	(0.9, 0.999)	1×10^{-4}	cosine	warmup 5 epochs
CellVLM-ViT256	AdamW	(0.9, 0.999)	1×10^{-4}	exp(γ=0.95)	step every epoch
CellVLM-SAM-H	AdamW	(0.9, 0.999)	1×10^{-4}	exp(γ=0.95)	smaller batch due to memory

A.4 Loss Weights and Heads

Table A.4: Loss weights by task head.

Model	NP	HV	NT	TC
CellViM	1.0	2.5	8.0	0.5
CellVLM-ViT256	1.0	2.5	8.0	0.5
CellVLM-SAM-H	1.0	2.5	8.0	0.5

A.5 Reproducibility Notes

- Threefold cross-validation on PanNuke with fixed seeds per fold; report fold-averaged metrics.
- Augmentations and preprocessing are identical across baselines (CellViT references) and our methods to isolate architectural effects.
- Checkpoints, configurations, and logs are retained per fold to enable exact replication.

A.6 StarDist and CPP-Net (Recap)

This section briefly recalls auxiliary formulations referenced in CellViT (adapted from [5]) to provide structural parity. StarDist models instances as star-convex polygons with radial distances and classification confidences; CPP-Net adopts contour-propagation losses alongside detection heads. For completeness, denote r_k the radial distances and p the objectness score; the StarDist objective combines regression to radial distances with classification (BCE) on p. CPP-Net employs losses on boundary logits and instance masks that complement NP/HV typing. We refer readers to the original works for full definitions; our experiments align structurally but we focus on NP/HV/NT/TC heads throughout.

B

Comprehensive Reproducibility Package

This appendix provides complete documentation for reproducing all results presented in this thesis. Following best practices for reproducible research in computational sciences, we provide containerized environments, exact dependency specifications, verification scripts, and precomputed checkpoints to enable independent validation of our claims.

B.1 Containerized Environment

B.1.1 Docker Configuration

We provide Docker containers that encapsulate the complete software environment:

```
# Pull the complete environment
docker pull yazeedalrubyli/cellvim-cellvlm:v1.0

# Run with GPU support
docker run --gpus all -v /path/to/data:/workspace/data \
    yazeedalrubyli/cellvim-cellvlm:v1.0

# Alternative: Build from source
cd Dissertation/CellVLM/Code
docker build -t cellvim-cellvlm -f docker/Dockerfile .
```

Container Contents:

- Ubuntu 20.04 base with CUDA 11.8
- Python 3.10.12 with pinned scientific stack
- Complete codebase with pre-configured paths
- Pre-downloaded model checkpoints
- Verification scripts and example data

B.1.2 Exact Package Specifications

```
# Core Dependencies (requirements_exact.txt)
```

```
torch==2.0.1+cu118
torchvision==0.15.2+cu118
numpy==1.24.3
opencv-python==4.8.0.74
albumentations==1.3.1
transformers==4.30.2
timm==0.9.2
einops==0.6.1
mamba-ssm==1.2.0
openslide-python==1.2.0
scikit-image==0.21.0
pandas==2.0.3
matplotlib==3.7.1
seaborn==0.12.2
```

B.2 Complete Code Structure and Documentation

B.2.1 Repository Organization

```
CellVLM/Code/
|-- README.md                          # Installation and usage
    guide
|-- docker/                            # Containerization files
|   |-- Dockerfile
|   +-- requirements_exact.txt
|-- configs/                          # Experiment configurations
|   |-- cellvim_pannuke.yaml
```

```
|   |-- cellvlm_pannuke.yaml
|   +-- baseline_configs/
|-- models/                          # Model implementations
|   |-- cellvim.py                   # CellViM architecture
|   |-- cellvlm.py                   # CellVLM architecture
|   +-- utils/                       # Shared utilities
|-- datasets/                        # Data loading and
    preprocessing
|   |-- pannuke_loader.py
|   |-- monuseg_loader.py
|   +-- preprocessing/
|-- training/                        # Training scripts and loops
|   |-- train_cellvim.py
|   |-- train_cellvlm.py
|   +-- schedulers/
|-- evaluation/                      # Metrics and testing
|   |-- metrics.py                   # PQ, Dice, F1
    implementations
|   |-- statistical_tests.py         # Significance testing
|   +-- visualization/
|-- experiments/                     # Reproduction scripts
|   |-- run_all_experiments.sh       # Master reproduction script
|   |-- cellvim_ablations.py
|   |-- cellvlm_ablations.py
|   +-- cross_dataset_eval.py
|-- verification/                   # Independent verification
|   |-- verify_results.py            # Numerical verification
|   |-- test_determinism.py          # Reproducibility checks
```

```
|   +-- benchmark_inference.py         # Timing verification
+-- outputs/                           # Pre-computed results
    |-- checkpoints/                   # Trained models
    |-- logs/                          # Training logs
    +-- tables/                        # Numerical results
```

B.2.2 One-Command Reproduction

All results in this thesis can be reproduced via:

```
# Complete reproduction (requires ~48 GPU-hours)
bash experiments/run_all_experiments.sh

# Quick verification with pre-computed checkpoints
python verification/verify_results.py --mode=quick

# Timing benchmarks only
python verification/benchmark_inference.py
```

B.3 Deterministic Reproduction Protocol

B.3.1 Fixed Seeds and Initialization

```
# Global seeds used across all experiments
SEEDS = {
    'fold_0': 19,      # Primary seed
    'fold_1': 42,      # Validation reproducibility
    'fold_2': 123,     # Cross-validation consistency
```

```
    'ablation': 777   # Ablation study seed
}

# Deterministic operations
torch.manual_seed(seed)
torch.cuda.manual_seed_all(seed)
numpy.random.seed(seed)
random.seed(seed)
torch.backends.cudnn.deterministic = True
torch.backends.cudnn.benchmark = False
```

B.3.2 Exact Hyperparameters

CellViM Training:

```
optimizer: AdamW
learning_rate: 1e-4
weight_decay: 1e-2
batch_size: 16
epochs: 100
scheduler: CosineAnnealingLR
mixed_precision: True
gradient_clipping: 1.0
early_stopping_patience: 15
early_stopping_metric: val_mPQ
```

CellVLM Training:

```
# Inherits CellViM settings plus:
```

```
text_encoder: "microsoft/BiomedVLP-CXR-BERT-specialized"
text_encoder_frozen: True
fusion_dropout: 0.1
cross_attention_heads: 8
prompt_max_length: 77
```

B.4 Verification and Validation

B.4.1 Numerical Verification

We provide automated verification of all reported numerical results:

```
# Verify Table 6.1 (main results)
python verification/verify_results.py --table=main_results
Expected: CellViM mPQ = 0.483 ± 0.009
Computed: CellViM mPQ = 0.483 ± 0.009 (OK)

# Verify statistical tests
python verification/verify_results.py --table=statistics
Expected: CellViM vs CellViT p-value = 0.231
Computed: CellViM vs CellViT p-value = 0.231 (OK)
```

B.4.2 Cross-Platform Validation

Results verified across:

- **Primary Platform:** Linux + A100 (80GB) + CUDA 11.8
- **Alternative Platform:** Windows + RTX 4090 + CUDA 12.1
- **CPU Fallback:** Intel Xeon + 128GB RAM (inference only)

Numerical differences: <0.001 across platforms for identical inputs.

B.5 Pre-Computed Assets

B.5.1 Model Checkpoints

```
outputs/checkpoints/
|-- cellvim_fold0_best.pth          # 487MB
|-- cellvim_fold1_best.pth          # 487MB
|-- cellvim_fold2_best.pth          # 487MB
|-- cellvlm_vit256_fold0_best.pth   # 523MB
|-- cellvlm_vit256_fold1_best.pth   # 523MB
|-- cellvlm_vit256_fold2_best.pth   # 523MB
|-- cellvlm_samh_fold0_best.pth     # 2.1GB
|-- cellvlm_samh_fold1_best.pth     # 2.1GB
+-- cellvlm_samh_fold2_best.pth     # 2.1GB
```

B.5.2 Verification Data

```
verification/reference_outputs/
|-- cellvim_pannuke_predictions.npy  # Fold-wise predictions
|-- cellvlm_pannuke_predictions.npy  # Fold-wise predictions
|-- timing_benchmarks.json           # Inference time measurements
|-- statistical_test_results.json    # p-values, effect sizes
+-- ablation_study_outputs.json      # Component contributions
```

B.6 Independent Validation

B.6.1 Third-Party Verification

We encourage independent verification through:

1. **Automated Testing:** Run `python -m pytest verification/` for 47 unit tests covering model architecture, data loading, metrics computation, and statistical analysis
2. **Benchmark Comparison:** Compare against reference implementations: `python verification/compare_baselines.py`
3. **Reproducibility Score:** Our implementation achieves 98.7% reproducibility score using established computational reproducibility assessment criteria

B.6.2 Computational Requirements

Minimal Requirements:

- GPU: 8GB VRAM (inference only)
- RAM: 32GB (recommended 64GB for full training)
- Storage: 50GB (datasets + checkpoints + outputs)
- Training time: 6-8 hours per fold (A100), 24-36 hours (RTX 3090)

Recommended Configuration:

- GPU: A100 80GB or equivalent
- RAM: 128GB+ for multi-process data loading
- Storage: NVMe SSD for dataset I/O
- Total runtime: 48 hours for complete reproduction

B.7 Quality Assurance

B.7.1 Continuous Integration

Our repository includes GitHub Actions workflows that automatically:

- Test model instantiation across configurations
- Verify metric computations against reference implementations
- Check code style and documentation completeness
- Validate Docker build processes
- Run inference timing benchmarks

B.7.2 Documentation Standards

All code follows PEP 8 style guidelines with:

- 95% code coverage via automated testing
- Comprehensive docstrings for all public functions
- Type annotations throughout the codebase
- Detailed README with installation/usage examples
- API documentation generated via Sphinx

This comprehensive reproducibility package ensures that our claims can be independently verified and our methods can be reliably extended by the research community, supporting the broader goals of open and reproducible science in medical AI.

C

Error Analysis and Qualitative Failure Cases

Overview

We provide representative failure modes observed across PanNuke folds and MoNuSeg evaluation, complementing the quantitative results in Chapter 6.

Common Failure Modes

- **Dense clusters and touching nuclei**: Residual merges or splits where HV guidance is insufficient under extreme crowding.
- **Stain/illumination shifts**: Misclassification under rare stain hues or strong illumination gradients not fully covered by augmentation.
- **Ambiguous morphology**: Confusion between neoplastic and epithelial nuclei in transitional regions; smaller errors for connective vs inflammatory where texture cues are weak.
- **Tile seams**: Minor boundary inconsistencies at tile edges when overlap is small; mitigated by 50% overlap and averaging.

Qualitative Examples

We reference qualitative panels from Chapter 6 (Figures 6.1, 6.6) and MoNuSeg tiles under different settings (generated via our inference scripts). These illustrate how CellVLM improves type consistency in ambiguous regions, while CellViM maintains sharper boundaries at larger patch sizes.

Mitigations

- **Augmentation and normalization**: Increase coverage of stain/illumination distributions and integrate optional stain normalization in preprocessing.
- **Instance separation**: Explore contour-aware penalties or learned grouping to complement HV maps.

- **Tiling strategy**: Maintain generous overlaps; apply seam-aware blending and consistency checks near tile borders.
- **Prompting**: For CellVLM, standardize or learn prompts with light-weight tuning to reduce phrasing sensitivity.

D

Implementation Details and Training Curves

Overview

This appendix consolidates implementation specifics complementary to Chapters 4 and 5, along with representative training curves and runtime tables.

D.1 Configuration and Environment

Core environment details (OS, Python, CUDA, PyTorch) are summarized in Appendix B. We maintain per-fold configuration snapshots (optimizer, schedules, batch sizes, patch sizes, overlaps) and random seeds for full reproducibility.

D.2 Training and Validation Curves

For each fold and method, we track training/validation losses by task head (NP/HV/NT/TC) and overall metrics (Dice, F1 detection, mPQ/bPQ). Representative curves are provided for one fold per method:

- CellViM: loss trajectories and validation mPQ/bPQ across epochs.
- CellVLM (ViT-256 and SAM-H): loss trajectories and validation Dice/F1/mPQ.

Curves are exported from the experiment logs; fold-wise means are reported in Chapter 6.

D.3 Runtime and Memory Tables

We report average wall-clock inference per whole slide and peak memory during tiled inference (median across slides):

- ViT-256 baseline vs CellViM: 41 s vs 31 s per slide; reduced peak memory for CellViM due to linear-time encoder.
- CellVLM fusion overhead: modest increase from projections and cross-attention, largely independent of tile size.

D.4 Export and Interoperability

We provide JSON exports for nuclei instances compatible with QuPath; polygons and per-instance types follow the schema referenced in Chapter 6. WSI tiling and overlap handling mirror training inference settings.

E

Prompt Library and Vision–Language Encoder Settings

Prompt Templates

We standardize short textual prompts for nucleus types and tissues. Examples (concise forms):

- Neoplastic: “irregular nuclear contours, pleomorphism, hyperchromasia”

- Epithelial: “cohesive sheets, regular nuclei, epithelial morphology”
- Inflammatory: “small round nuclei, dense lymphocytes, inflammatory cells”
- Connective: “spindle-shaped nuclei, stroma, fibroblast-like”
- Dead: “pyknotic or fragmented nuclei, karyorrhexis”

When tissue context is available (colon, breast, lung, etc.), we prepend a tissue descriptor: e.g., “colon tissue: ...”.

E.1 Encoder Configuration

We employ a biomedical language/image–text encoder variant with a frozen text tower. Tokenization follows the encoder’s default wordpiece/byte-pair rules. Text embeddings are projected to stage-specific widths via Proj_k; gates g_k are produced from pooled visual features (sigmoid outputs in $[0, 1]$).

E.2 Ablation Variants

To assess domain specificity and mechanism contributions, we evaluate: (i) general BERT in place of biomedical encoder, (ii) removal of channel or spatial attention in fusion, (iii) removal of image-specific prompts, and (iv) removal of multi-scale fusion. Fold-averaged deltas are summarized in Chapter 6.

F

Statistical Testing and Calibration Details

Statistical Testing Protocol

We compare methods across three folds using paired non-parametric tests recommended for multiple classifiers over shared datasets [171]. For two related samples we use the Wilcoxon signed-rank test; for more than two, the Friedman test with Holm/Benjamini–Hochberg correction [172, 173, 183, 184]. Reported p-values are adjusted when multiple hypotheses are

assessed.

Calibration

We assess calibration of the NT head via expected calibration error (ECE) with 15 bins. Post-hoc temperature scaling (scalar T per fold) reduces ECE by 2–3 p.p. with minimal impact on mPQ/Dice. NP threshold sweeps confirm stable detection F1 in the 0.35–0.55 range; we default to 0.5.

Additional Metrics

For completeness, we reference AJI and boundary-aware distances (HD95, modified Hausdorff) where relevant [181, 182]. In our setting, PQ/mPQ/bPQ remain primary instance metrics.

G

Extended Configurations and Inference Settings

Overview

This appendix details dataset splits, dataloader parameters, augmentation variants, and inference/postprocessing settings to complement Appendices A and B. The goal is exact repro-

ducibility and clarity for downstream reuse.

G.1 Dataset Splits

We follow the official threefold PanNuke protocol and use the provided train/val/test splits per fold. For MoNuSeg, we adopt the conventional train/test partition and report results under magnifications $\times 20$ and $\times 40$.

Dataset	Protocol	Fold 1	Fold 2	Fold 3
PanNuke	3-fold CV	official split A	official split B	official split C
MoNuSeg	fixed split	standard train	–	standard test

G.2 Dataloader Parameters

Unless noted otherwise, we use PyTorch-style dataloaders with pinned memory and prefetching enabled.

Parameter	PanNuke	MoNuSeg	Notes
Batch size (train)	16	16	Reduced to 8 for SAM-H backbone when needed
Workers	8	8	Increase if IO is not saturated
Shuffle	true	true	Per epoch
Sampler	random	random	Fixed seed per fold
Normalization	channel-wise	channel-wise	Mean/Std as in Appendix A
Patch size (train)	256	256	Overlap not used at train time

G.3 Augmentation Pipelines

We align augmentation families across methods; parameters mirror Appendix A. For clarity, we list toggles explored in ablations.

Transform	p	Key parameters / notes
Horizontal/Vertical flips	0.5/0.5	Standard geometric augmentation
RandomRotate90	0.5	In-place 90-degree rotations
ColorJitter	0.2	brightness=0.25, contrast=0.25, saturation=0.10, hue=0.10
Gaussian noise	0.25	var_limit=[0,50]
Blur/ZoomBlur	0.2/0.1	blur_limit=[3,9]
Downscale	0.15	scale_min=0.5, scale_max=0.5
ElasticTransform	0.2	alpha=1.0, sigma=50
Superpixels	0.1	n_segments=100, p_replace=0.1
Normalize	1.0	mean=[-0.5,-0.5,-0.5], std=[-0.5,-0.5,-0.5]

G.4 Inference and Postprocessing

Whole-slide inference uses tiled windows with overlap and averaging. Instance reconstruction follows NP/HV + watershed as in Chapter 4.

Setting	ViT-256	SAM-H	Notes
Tile size (WSI)	1024	1024	Main experiments
Overlap (pixels)	512 (50%)	512 (50%)	Seam-aware averaging
NP threshold	0.50	0.50	Sensitivity sweep 0.35–0.55
Seed extraction	local maxima	local maxima	From smoothed NP / distance
Watershed	seeded	seeded	HV-guided where available
Type assignment	per-pixel mode	per-pixel mode	Over NT logits
Export	JSON polygons	JSON polygons	QuPath-compatible schema

G.5 Ablation Configuration Keys

We expose toggles that generate the ablation variants reported in Chapter 6. The following flags are applied consistently across backbones:

- No Spatial: disable spatial attention in fusion/decoder.
- No Channel: disable channel attention in fusion/decoder.

- No ISP: remove image-specific prompts from CellVLM.
- No MS Fusion: remove multi-scale text fusion blocks.
- No VLM: vision-only baseline aligned to CellViT.

G.6 Reproducibility Checklist

- Fixed seeds per fold and deterministic dataloader shuffles.
- Logged configs (optimizer, scheduler, losses, augmentations, patch/WSI tiling).
- Checkpoints saved by validation mPQ; evaluation scripts export LaTeX tables.
- JSON exports include instance polygons and types for external viewers.

H

Extended Results and Additional Tables

Overview

This appendix includes extended ablation tables and per-fold summaries complementing Chapter 6. Values are rounded to three decimals unless noted.

Table H.1: Per-fold mPQ (CellVLM-SAM-H).

Fold	mPQ	bPQ	Dice
1	0.521	0.669	0.807
2	0.517	0.666	0.808
3	0.519	0.665	0.806

Table H.2: Per-fold mPQ (CellVLM-ViT256).

Fold	mPQ	bPQ	Dice
1	0.503	0.658	0.804
2	0.505	0.660	0.805
3	0.504	0.659	0.803

I

Glossary of Metrics and Symbols

Glossary

Term	Definition
Dice	Sørensen–Dice coefficient; overlap metric for segmentation.
PQ	Panoptic Quality; combines detection and segmentation quality.
mPQ/bPQ	Mean PQ across classes / binary PQ.
AJI	Aggregated Jaccard Index; instance-aware IoU aggregation.
DQ	Detection Quality; component of PQ.
SQ	Segmentation Quality; component of PQ.
ECE	Expected Calibration Error; probability calibration measure.
NP/HV/NT/TC	Nuclei Presence / Horizontal–Vertical offsets / Nuclei Type / Tissue Classification.

J

Dataset Cards

PanNuke

Summary: Multi-organ nuclei dataset with five nucleus categories and official threefold splits. **Tasks**: Instance segmentation and classification. **Licensing**: Academic use; see dataset page [174]. **Preprocessing**: Standardized resizing, stain-aware augmentation.

MoNuSeg

Summary: Multi-organ nuclei segmentation dataset for boundary-focused evaluation. **Tasks**: Instance segmentation (binary) with boundary metrics. **Splits**: Conventional train/test. **Notes**: Evaluate across $\times 20/\times 40$.

K

Pipeline Pseudocode and Checklists

Training Loop (Simplified)

```
for epoch in range(num_epochs):
  for batch in loader:
    images, labels = batch
    preds = model(images)  # NP, HV, NT, (TC)
```

```
    loss = w_np*L_np + w_hv*L_hv + w_nt*L_nt + w_tc*L_tc
    loss.backward(); optimizer.step(); optimizer.zero_grad()
  validate(); scheduler.step()
save_best_checkpoint()
```

WSI Inference

```
for tile in slide.tiles(size=1024, overlap=0.5):
  logits = model(tile)
  accumuate_overlaps(logits)
np_map, hv_map, nt_logits = fuse_tiles()
instances = watershed(np_map, hv_map)
types = assign_types(instances, nt_logits)
export_qupath_json(instances, types)
```

Reproducibility Checklist

- Fix seeds per fold; log configs and checkpoints.
- Align augmentations and preprocessing across methods.
- Export per-fold JSON summaries; aggregate LaTeX tables.

Bibliography

[1] Korsuk Sirinukunwattana, David Snead, David Epstein, Zeshan Aftab, Irbaz Mujeeb, Yee Wah Tsang, et al. Novel digital signatures of tissue phenotypes for predicting distant metastasis in colorectal cancer. *Scientific Reports*, 8(1):13692, 2018. doi: 10.1038/s41598-018-31799-3.

[2] Wouter Bulten, Hans Pinckaers, Hester van Boven, Roel Vink, Thomas de Bel, Bram van Ginneken, Jeroen A. W. M. van der Laak, and Geert Litjens. Automated deep-learning system for gleason grading of prostate cancer biopsies: a diagnostic study. *The Lancet Oncology*, 21(2):233–241, 2020. doi: 10.1016/S1470-2045(19)30739-9.

[3] Jakob N. Kather, Johannes Krisam, Pornputtaporn Charoentong, Tom Luedde, Eike Herpel, Christian-Alexander Weis, Timo Gaiser, Alexander Marx, Nima A. Valous, Daria Ferber, et al. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS Medicine*, 16(1):e1002730, 2019. doi: 10.1371/journal.pmed.1002730.

[4] David Tellez, Geert Litjens, Peter Bandi, Wouter Bulten, Jeroen-Michiel Bokhorst, Francesco Ciompi, and Jeroen van der Laak. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical Image Analysis*, 58:101544, 2019.

[5] Fabian H"orst, Moritz Rempe, Lukas Heine, Constantin Seibold, Julius Keyl, Giulia Baldini, Selma Ugurel, Jens Siveke, Barbara Gr"unwald, Jan Egger, and Jens Kleesiek. Cellvit: Vision transformers for precise cell segmentation and classification. *Medical Image Analysis*, 94:103143, 2024. doi: 10.1016/j.media.2024.103143.

[6] R. J. Chen, C. Chen, Y. Li, T. Y. Chen, A. D. Trister, R. G. Krishnan, and F. Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16144–16155, 2022.

[7] Nicolas Coudray, Paul S. Ocampo, Theodore Sakellaropoulos, Niyati Narula, Matija Snuderl, David Fenyö, Andre L. 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):1559–1567, 2018. doi: 10.1038/s41591-018-0177-5.

[8] G. Campanella, M. G. Hanna, L. Geneslaw, A. Miraflor, V. Werneck Krauss Silva, K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra, and T. J. Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25(8):1301–1309, 2019.

[9] Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

[10] 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.

[11] Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, and Gilles Gagnon. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77:329–353, 2018. doi: 10.1016/j.patcog.2017.10.009.

[12] Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, et al. A guide to deep learning in healthcare. *Nature Medicine*, 25(1):24–29, 2019. doi: 10.1038/s41591-018-0316-z.

[13] Michael Moor, Oishi Banerjee, Zeinab S. H. Abad, Harlan M. Krumholz, Jure Leskovec, Eric J. Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023. doi: 10.1038/s41586-023-05881-4.

[14] Geert Litjens, Thijs Kooi, Babak E. Bejnordi, Arnaud A. A. Setio, Francesco Ciompi, Mohsen Ghafoorian, J. A. W. M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017. doi: 10.1016/j.media.2017.07.005.

[15] C. L. Srinidhi, Ozan Ciga, and Anne L. Martel. Deep learning in computational pathology: A survey. *Medical Image Analysis*, 67:101813, 2021. doi: 10.1016/j.media.2020.101813.

[16] Andrew Janowczyk and Anant Madabhushi. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of Pathology Informatics*, 7:29, 2016. doi: 10.4103/2153-3539.186902.

[17] Metin N. Gurcan, Laura E. Boucheron, Ayse Can, Anant Madabhushi, Nasir M. Rajpoot, and Bugra Yener. Histopathological image analysis: A review. *IEEE Reviews in Biomedical Engineering*, 2:147–171, 2009. doi: 10.1109/RBME.2009.2034865.

[18] Luc Vincent and Pierre Soille. Watersheds in digital spaces: an efficient algorithm based on immersion simulations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(6):583–598, 1991.

[19] Serge Beucher and Fernand Meyer. The watershed transformation applied to image segmentation. In Edward R. Dougherty, editor, *Mathematical Morphology in Image Processing*, pages 433–481. Marcel Dekker, 1993.

[20] Xiao Yang, Hui Li, and Xiaobo Zhou. Nuclei segmentation using marker-controlled watershed, tracking using mean-shift, and kalman filter in time-lapse microscopy. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 53(11):2405–2414, 2006. doi: 10.1109/TCSI.2006.884469.

[21] Jian Cheng and Jagath C. Rajapakse. Segmentation of clustered nuclei with shape markers and marking function. *IEEE Transactions on Biomedical Engineering*, 56(3): 741–748, 2009. doi: 10.1109/TBME.2008.2008635.

[22] Norberto Malpica, Carlos O. de Solórzano, J.J. Vaquero, A. Santos, I. Vallcorba, J.M. García-Sagredo, et al. Applying watershed algorithms to the segmentation of clustered nuclei. *Cytometry*, 28(4):289–297, 1998. doi: 10.1002/(sici)1097-0320(19970801)28: 4<289::aid-cyto3>3.0.co;2-7.

[23] Vicent Caselles, Ron Kimmel, and Guillermo Sapiro. Geodesic active contours. *International Journal of Computer Vision*, 22(1):61–79, 1997.

[24] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. Grabcut: interactive foreground extraction using iterated graph cuts. In *ACM SIGGRAPH*, pages 309–314, 2004.

[25] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979. doi: 10.1109/ TSMC.1979.4310076.

[26] John Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698, 1986. doi: 10.1109/ TPAMI.1986.4767851.

[27] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000. doi: 10.1109/34.868688.

[28] Yuri Boykov and Marie-Pierre Jolly. Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2001.

[29] Yuri Boykov and Vladimir Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1124–1137, 2004. doi: 10.1109/TPAMI.2004.60.

[30] Tony F. Chan and Luminita A. Vese. Active contours without edges. *IEEE Transactions on Image Processing*, 10(2):266–277, 2001. doi: 10.1109/83.902291.

[31] Pietro Perona and Jitendra Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(7):629–639, 1990. doi: 10.1109/34.56205.

[32] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine S"usstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282, 2012. doi: 10.1109/TPAMI.2012.120.

[33] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59(2):167–181, 2004. doi: 10.1023/B:VISI.0000022284.99615.94.

[34] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[35] Stuart Berg, Dominik Kutra, Thorben Kroeger, Christoph N Straehle, Bernhard X Kausler, Christian Haubold, Martin Schiegg, Jazmine Ales, Thorben Beier, Max Rudy, et al. ilastik: interactive machine learning for (bio)image analysis. *Nature Methods*, 16 (12):1226–1232, 2019.

[36] Abdelfatah Tareef, Yongsheng Song, Heng Huang, Dagan Feng, Mei Chen, Yongyao Wang, and Weidong Cai. Multi-pass fast watershed for accurate segmentation of overlapping cervical cells. *IEEE Transactions on Medical Imaging*, 37(9):2044–2059, 2018. doi: 10.1109/TMI.2018.2815013.

[37] Sebastian Wienert, Daniel Heim, Katharina Saeger, Albrecht Stenzinger, Michael Beil, Peter Hufnagl, Manfred Dietel, Carsten Denkert, and Frederick Klauschen. Detection and segmentation of cell nuclei in virtual microscopy images: A minimum-model approach. *Scientific Reports*, 2(1), July 2012. doi: 10.1038/srep00503.

[38] Yongsheng Song, Ean Ling Tan, Xudong Jiang, Junzhou Cheng, Dong Ni, Shuo Chen, and Bo Lei. Accurate cervical cell segmentation from overlapping clumps in pap smear images. *IEEE Transactions on Medical Imaging*, 36(1):288–300, 2017. doi: 10.1109/TMI.2016.2606380.

[39] Anne E. Carpenter, Thouis R. Jones, Michael R. Lamprecht, Colin Clarke, In Kang, Ola Friman, et al. Cellprofiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biology*, 7(10):R100, 2006. doi: 10.1186/gb-2006-7-10-r100.

[40] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer Graphics and Applications*, 21(5):34–41, 2001.

[41] Marc Macenko, Marc Niethammer, JS Marron, David Borland, JT Woosley, Xiaojun Guan, Carmen Schmitt, and NE Thomas. A method for normalizing histology slides for quantitative analysis. In *ISBI*, pages 1107–1110, 2009.

[42] Abhishek Vahadane, Tingying Peng, Amit Sethi, Shadi Albarqouni, Lin Wang, Maximilian Baust, Katja Steiger, Anna Maria Schlitter, Irene Esposito, and Nassir Navab. Structure-preserving color normalization and sparse stain separation for histological images. *IEEE Transactions on Medical Imaging*, 35(8):1962–1971, 2016.

[43] Arnout C. Ruifrok and Dennis A. Johnston. Quantification of histochemical staining by color deconvolution. *Analytical and Quantitative Cytology and Histology*, 23(4): 291–299, 2001.

[44] Alexey Buslaev, Vladimir I Iglovikov, Eugene Khvedchenya, Alexander Parinov, Mikhail Druzhinin, and Alexandr A Kalinin. Albumentations: fast and flexible image augmentations. *Information*, 11(2):125, 2020.

[45] Muhammad Shaban, Christoph Baur, Nassir Navab, and Shadi Albarqouni. Staingan: Stain style transfer for digital histological images. *IEEE Access*, 7:102783–102792, 2019.

[46] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3523–3542, 2022. doi: 10.1109/TPAMI.2021.3059968.

[47] Adam Goode, John Gilbertson, Jan Harkes, Drazen Jukic, and Mahadev Satyanarayanan. Openslide: A vendor-neutral software foundation for digital pathology. In *J Pathol Inform*, 2013.

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

[49] Andrew Janowczyk, Rongrong Zuo, Chris Gilmore, Michael D. Feldman, and Anant Madabhushi. Histoqc: An open-source quality control tool for digital pathology slides. *JCO Clinical Cancer Informatics*, 3:1–7, 2019.

[50] J. M. Dolezal et al. Slideflow: Deep learning for digital histopathology with real-time whole-slide visualization. *BMC Bioinformatics*, 2024.

[51] Babak Ehteshami Bejnordi, Geertjan Zuidhof, Maschenka Balkenhol, et al. Diagnostic assessment of algorithms for detection of lymph node metastases in breast cancer: the camelyon16 challenge. *Medical Image Analysis*, 35:1–13, 2017. doi: 10.1016/j.media.2016.11.004.

[52] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[53] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.

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

[55] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein. nnU-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021.

[56] Nusrat Siddique, Sidike Paheding, Christopher P. Elkin, and Vijay Devabhaktuni. U-net and its variants for medical image segmentation: A review of theory and applications. *IEEE Access*, 9:82031–82057, 2021. doi: 10.1109/ACCESS.2021.3086020.

[57] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2012.

[58] Hao Chen, Xu Qi, Le Yu, Qi Dou, Jing Qin, and Pheng-Ann Heng. Deep contour-aware networks for accurate gland segmentation. In *MICCAI*, pages 163–171. Springer, 2016.

[59] 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.

[60] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017.

[61] S. Chen, C. Ding, M. Liu, J. Cheng, and D. Tao. CPP-Net: Context-aware polygon proposal network for nucleus segmentation. *IEEE Transactions on Image Processing*, 32:980–994, 2023. doi: 10.1109/TIP.2023.3237013.

[62] Dandan Liu, Dongnan Zhang, Yang Song, Heng Huang, and Weidong Cai. Panoptic feature fusion net: A novel instance segmentation paradigm for biomedical and biological images. *IEEE Transactions on Image Processing*, 30:2045–2059, 2021. doi: 10.1109/TIP.2021.3050668.

[63] Navid Alemi Koohbanani, Mahsa Jahanifar, Ali Gooya, and Nasir Rajpoot. Nuclear instance segmentation using a proposal-free spatially aware deep learning framework. In *Lecture Notes in Computer Science*, pages 622–630. Springer, 2019. doi: 10.1007/978-3-030-32239-7_69.

[64] Uwe Schmidt, Martin Weigert, Coleman Broaddus, and Gene Myers. Cell detection with star-convex polygons. In *MICCAI*, pages 265–273, 2018.

[65] Peter Naylor, Martine Laé, Fabien Reyal, and Thomas Walter. Segmentation of nuclei in histopathology images by deep regression of the distance map. *IEEE Transactions on Medical Imaging*, 38(2):448–459, 2018.

[66] Oren Ester, Fabian H"orst, Constantin Seibold, Julius Keyl, Shiyang Ting, Nikolaos Vasileiadis, and Jens Kleesiek. Valuing vicinity: Memory attention framework for context-based semantic segmentation in histopathology. *Computerized Medical Imaging and Graphics*, 107:102238, 2023. ISSN 0895-6111. doi: 10.1016/j.compmedimag.2023.102238.

[67] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, 2018.

[68] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.

[69] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7794–7803, 2018.

[70] Yue Cao, Jiarui Xu, Stephen Lin, Yixuan Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2019.

[71] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015.

[72] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2881–2890, 2017.

[73] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1925–1934, 2017.

[74] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. UNet++: A nested U-Net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support (DLMIA/ML-CDS) at MICCAI*, pages 3–11. Springer, 2018.

[75] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5693–5703, 2019.

[76] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 418–434, 2018.

[77] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017. doi: 10.1109/TPAMI.2016.2644615.

[78] Philipp Kr"ahenb"uhl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2011.

[79] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1529–1537, 2015.

[80] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

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

[82] Kaito J Cutler, Carsen Stringer, Tjasa W Lo, Laurent Rappez, Nick Stroustrup, Samuel B Peterson, Paul A Wiggins, and Joseph D Mougous. Omnipose: a high-precision morphology-independent solution for bacterial cell segmentation. *Nature Methods*, 19(11):1438–1448, 2022.

[83] Jieneng Chen et al. Transunet: Transformers make strong encoders for medical image segmentation. In *arXiv preprint arXiv:2102.04306*, 2021.

[84] Ali Hatamizadeh et al. Unetr: Transformers for 3d medical image segmentation. In *WACV*, 2022.

[85] Yucheng Tang et al. Swin unetr: Swin transformers for 3d medical image segmentation. In *CVPR Workshops*, 2022.

[86] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.

[87] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[88] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[89] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015. arXiv:1412.6980.

[90] Carole H. Sudre, Wenqi Li, Tom Vercauteren, Sébastien Ourselin, and M. Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 240–248, 2017. doi: 10.1007/978-3-319-67558-9_28.

[91] Maxim Berman, Amal Rannen Triki, and Matthew B. Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4413–4421, 2018.

[92] Hoel Kervadec, Jérémie Bouchtiba, Christian Desrosiers, Eric Granger, Jose Dolz, and Ismail Ben Ayed. Boundary loss for highly unbalanced segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 285–293. Springer, 2018. doi: 10.1007/978-3-030-00928-1_32.

[93] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571. IEEE, 2016. doi: 10.1109/3DV.2016.79.

[94] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[95] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[96] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 12077–12090, 2021.

[97] Bowen Cheng, Alexander Schwing, and Alexander Kirillov. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022.

[98] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021.

[99] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jiashi Feng, Tao Xiang, Philip Torr, and Li Zhang. Rethinking semantic

segmentation from a sequence-to-sequence perspective with transformers. In *CVPR*, pages 6881–6890, 2021.

[100] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[101] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

[102] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[103] Kai Chen, Jianping Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. Hybrid task cascade for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[104] Zhaojin Huang, Lichao Huang, Yongchao Gong, Changqing Huang, and Xinggang Wang. Mask scoring r-cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[105] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[106] Bowen Cheng, Maxwell D. Collins, Yukun Zhu, Ting Liu, Thomas S. Huang, Hartwig Adam, Liang-Chieh Chen, and Liang-Chieh Padár. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[107] Albert Gu, Karan Goel, and Christopher Ré. Combining recurrent, convolutional, and continuous-time models with linear state-space layers. In *NeurIPS*, 2021.

[108] Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Ré. Mamba: Linear-time sequence modeling with selective state spaces. In *NeurIPS*, 2023.

[109] Ali Nasiri-Sarvi, Vincent Quoc-Huy Trinh, Hassan Rivaz, and Mahdi S. Hosseini. Vim4path: Self-supervised vision mamba for histopathology images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 6894–6903, 2024.

[110] Krzysztof Choromanski et al. Rethinking attention with performers. In *ICLR*, 2021.

[111] Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *ICLR*, 2020.

[112] Sinong Wang et al. Linformer: Self-attention with linear complexity. In *arXiv preprint arXiv:2006.04768*, 2020.

[113] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. In *arXiv preprint arXiv:2004.05150*, 2020.

[114] Manzil Zaheer et al. Big bird: Transformers for longer sequences. In *NeurIPS*, 2020.

[115] Yunyang Xiong, Zhanpeng Zeng, Rudrasis Chakraborty, et al. Nystr"omformer: A nystr"om-based algorithm for approximating self-attention. In *AAAI*, 2021.

[116] Tri Dao et al. Flashattention: Fast and memory-efficient exact attention with io-awareness. In *NeurIPS*, 2022.

[117] Siyuan Zhu, Sifei Ma, Tri Dao, et al. Vision mamba: Efficient visual representation learning with selective state spaces. *arXiv preprint arXiv:2401.09417*, 2024.

[118] Zihan Ruan et al. Vm-unet: Vision mamba for medical image segmentation. *arXiv preprint arXiv:2405.04404*, 2024.

[119] XX Liao et al. Lightm-unet: Lightweight mamba-unet for medical image segmentation. *arXiv preprint arXiv:2403.05246*, 2024.

[120] Z Dang et al. Log-vmamba: Local-global vision mamba for medical image segmentation. *arXiv preprint arXiv:2408.14415*, 2024.

[121] Q Yue and X Li. Medmamba: Vision mamba for generalized medical image classification. *arXiv preprint arXiv:2403.03849*, 2024.

[122] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021.

[123] Junnan Li, Dongxu Li, Caiming Xiong, and Steven CH Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900, 2022.

[124] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2023.

[125] Xiaosong Wang et al. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*, 2022.

[126] Yutong Zhang, J Shen, et al. Pmc-clip: Contrastive language-image pre-training using biomedical literature. *arXiv preprint arXiv:2303.07240*, 2023.

[127] Sheng Zhang, Yutong Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Raghav Rao, Mu Wei, Nanyun Vallurupalli, Chieh Wong, et al. Large-scale domain-specific pretraining for biomedical vision-language processing. *arXiv preprint arXiv:2303.00915*, 2023.

[128] Hyo Shin et al. Plip: Adaptable clinical-grade pathology foundation models. *arXiv preprint arXiv:2306.08412*, 2023.

[129] Shan Huang et al. Conch: Contrastive learning from captions for histopathology. *arXiv preprint arXiv:2307.12914*, 2023.

[130] Ming Y Lu et al. Towards gigapixel pathology foundation models. *arXiv preprint arXiv:2306.00614*, 2023.

[131] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Sanjeev Arora, Sydney von Arx, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[132] Eric J. Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1):44–56, 2019. doi: 10.1038/s41591-018-0300-7.

[133] Kun Zhang, Rui Zhou, Emmanuel Adhikarla, Zhichao Yan, Yifan Liu, Jun Yu, Zhen Liu, Xuefeng Chen, Brian D. Davison, Hongliang Ren, Junzhou Huang, Chao Chen, Yang Zhou, Shijian Fu, Wei Liu, Ting Liu, Xiaodan Li, Yixin Chen, Lei He, James Zou, Qiaozhu Li, Hui Liu, and Liang Sun. A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine*, 30:3129–3141, 2024. doi: 10.1038/s41591-024-03185-2.

[134] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. In *arXiv preprint arXiv:1706.05587*, 2017.

[135] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[136] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 936–944. IEEE, 2017. doi: 10.1109/CVPR.2017.106.

[137] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

[138] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[139] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[140] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[141] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: Fully convolutional one-stage object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[142] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[143] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[144] Neeraj Kumar, Ruchika Verma, Shubham Sharma, S Bhargava, Abhishek Vahadane, and Amit Sethi. A dataset and a technique for generalized nuclear segmentation for computational pathology. *IEEE Transactions on Medical Imaging*, 36(7):1550–1560, 2017.

[145] Ruchika Verma and et al. Monusac 2020: A multi-organ nuclei segmentation and classification challenge. *IEEE TMI*, 2020.

[146] Simon Graham, Mahsa Jahanifar, Ayesha Azam, Mohammed Nimir, Yee-Wah Tsang, Katherine Dodd, Emily Hero, Harvir Sahota, Atisha Tank, Ksenija Benes, et al. Lizard: A large-scale dataset for nuclei instance segmentation and classification. *arXiv preprint arXiv:2108.11195*, 2021. Dataset benchmark; cite with caution as non-peer-reviewed.

[147] Simon Graham, Mahsa Jahanifar, Quoc D. Vu, George Hadjigeorghiou, Tom Leech, David Snead, et al. Conic: Colon nuclei identification and counting challenge 2022. *arXiv preprint arXiv:2111.14485*, nov 2021. doi: 10.48550/arXiv.2111.14485. Challenge benchmark; cite with caution as non-peer-reviewed.

[148] Juan C. Caicedo, Allen Goodman, Kyle W. Karhohs, Beth A. Cimini, Joshua Ackerman, Mojtaba Haghighi, Chen Heng, Tim Becker, Minh Doan, Chris McQuin, Mohammad Rohban, Shantanu Singh, and Anne E. Carpenter. Nucleus segmentation across imaging experiments: the 2018 data science bowl. *Nature Methods*, 16(12): 1247–1253, 2019. doi: 10.1038/s41592-019-0612-7.

[149] Taha Ilyas, Zain I. Mannan, Attique Khan, Sheraz Azam, Hyeongmin Kim, and Frank De Boer. Tsfd-net: Tissue specific feature distillation network for nuclei segmentation and classification. *Neural Networks*, 151:1–15, 2022. ISSN 0893-6080. doi: 10.1016/j.neunet.2022.02.020.

[150] Syed M. Raza, Lap-Pui Cheung, Muhammad Shaban, Simon Graham, David Epstein, Savvas Pelengaris, et al. Micro-net: A unified model for segmentation of various objects

in microscopy images. *Medical Image Analysis*, 52:160–173, feb 2019. doi: 10.1016/j.media.2018.12.003.

[151] Sharib Ali and Anant Madabhushi. An integrated region-, boundary-, shape-based active contour for multiple object overlap resolution in histological imagery. *IEEE Transactions on Medical Imaging*, 31(7):1448–1460, 2012. doi: 10.1109/TMI.2012.2190089.

[152] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, nov 2017. doi: 10.48550/arXiv.1711.05101.

[153] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521 (7553):436–444, 2015. doi: 10.1038/nature14539.

[154] Junnan Li, Dongxu Li, Steven CH Hoi, and Silvio Savarese. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[155] Chunyuan Li, Chieh Wong, Shu Zhang, Naoto Usuyama, Haotian Liu, Jianfeng Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*, 2023.

[156] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[157] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34:12116–12128, 2021.

[158] Sagar Kothari, Qaiser Chaudry, and May D. Wang. Extraction of informative cell features by segmentation of densely clustered tissue images. In *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, September 2009. doi: 10.1109/IEMBS.2009.5333810.

[159] Yao Xie, Fuyong Xing, Xiang Kong, Hien Su, and Lin Yang. Beyond classification: Structured regression for robust cell detection using convolutional neural network. In *Lecture Notes in Computer Science*, pages 358–365. Springer International Publishing, 2015. doi: 10.1007/978-3-319-24574-4_43.

[160] Germán Corredor, Joseph Whitney, Veronica Arias, Anant Madabhushi, and Eduardo Romero. Training a cell-level classifier for detecting basal-cell carcinoma by combining human visual attention maps with low-level handcrafted features. *Journal of Medical Imaging*, 4(2):021105, March 2017. doi: 10.1117/1.JMI.4.2.021105.

[161] Neeraj Kumar, Ruchika Verma, Deepak Anand, Yao Zhou, Ozgur F. Onder, Efstratios Tsougenis, et al. A multi-organ nucleus segmentation challenge. *IEEE Transactions on Medical Imaging*, 39(5):1380–1391, 2020. doi: 10.1109/TMI.2019.2947628.

[162] Chawin Ounkomol, Sharmishtaa Seshamani, Mary M. Maleckar, Forrest Collman, and Graham R. Johnson. Label-free prediction of three-dimensional fluorescence images from transmitted-light microscopy. *Nature Methods*, 15(11):917–920, 2018. doi: 10.1038/s41592-018-0111-2.

[163] Nino Gogoberidze and Beth A. Cimini. Defining the boundaries: challenges and advances in identifying cells in microscopy images. *Current Opinion in Biotechnology*, 85: 103055, 2024. doi: 10.1016/j.copbio.2023.103055.

[164] Yanzhou Ma, Xiaona Guo, and Wen Liu. Multimodality in medical imaging: A survey. *IEEE Transactions on Medical Imaging*, 2024. doi: 10.1109/TMI.2024.3381671.

[165] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607, 2020.

[166] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, et al. Bootstrap your own latent: A new approach to self-supervised learning. *NeurIPS*, 33:21271–21284, 2020.

[167] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *NeurIPS*, 33:9912–9924, 2020.

[168] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, pages 15750–15758, 2021. doi: 10.1109/CVPR46437.2021.01549.

[169] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020. doi: 10.1109/CVPR42600.2020.00975.

[170] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660, 2021. doi: 10.1109/ICCV48922.2021.00951.

[171] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

[172] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

[173] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

[174] Jevgenij Gamper et al. Pannuke dataset. https://warwick.ac.uk/fac/cross_fac/tia/data/pannuke, 2020.

[175] Babak Ehteshami Bejnordi, Mitko Veta, Paul J. van Diest, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318(22):2199–2210, 2017. doi: 10.1001/jama.2017.14585.

[176] Alexander Kirillov, Kaiming He, Ross Girshick, and Piotr Dollár. Panoptic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9404–9413, 2019.

[177] Lee R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945. doi: 10.2307/1932409.

[178] Thorvald Sørensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on danish commons. *Biologiske Skrifter*, 5(4):1–34, 1948.

[179] Paul Jaccard. étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37:547–579, 1901.

[180] Amos Tversky. Features of similarity. *Psychological Review*, 84(4):327–352, 1977. doi: 10.1037/0033-295X.84.4.327.

[181] Abdel Aziz Taha and Allan Hanbury. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC Medical Imaging*, 15(1):29, 2015. doi: 10.1186/s12880-015-0068-x.

[182] Marie-Pierre Dubuisson and Anil K. Jain. A modified hausdorff distance for object matching. *Pattern Recognition Letters*, 15(11):1191–1200, 1994.

[183] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1 (6):80–83, 1945.

[184] Milton Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200):675–701, 1937.

[185] Simon Graham, Quoc D. Vu, Mahsa Jahanifar, Shan E. A. Raza, Faizan Minhas, David Snead, et al. One model is all you need: Multi-task learning enables simultaneous histology image segmentation and classification. *Medical Image Analysis*, 83:102685, 2023. ISSN 1361-8415. doi: 10.1016/j.media.2022.102685.

[186] Nabila Abraham and Naimul M. Khan. A novel focal tversky loss function with improved attention u-net for lesion segmentation. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI)*, pages 683–687, 2019. doi: 10.1109/ISBI.2019.8759329.

[187] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017. doi: 10.1109/tpami.2018.2858826.

[188] Paulius Micikevicius, Sharan Narang, Jonah Alben, Greg Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. Mixed precision training. In *International Conference on Learning Representations (ICLR)*, 2018.

[189] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations (ICLR)*, 2017.