THE UNIVERSITY *of* EDINBURGH

Deep Learning Methods for the Automated Segmentation and Phenotyping of Cells in Biological Tissues

Thibaut Goldsborough

Doctor of Philosophy

The University of Edinburgh

2025

To Minimags,

Abstract

This thesis aims to advance the field of tissue phenotyping by developing and refining computational tools to accurately analyse and quantify microscopy images of biological tissues. Tissue phenotyping involves characterizing cellular composition, architecture, and function. This is essential for understanding the biological processes underpinning multiple diseases, including cancer. With the increasing complexity of microscopy data, it is crucial to rely on quantitative methods that can detect, identify, and characterize individual cells and their interactions.

The initial focus of this research is to improve the accuracy of cell and nucleus segmentation methods, as this is often a critical step in subsequent analyses. We introduce InstanSeg, a novel, embedding-based instance segmentation algorithm optimized for accuracy, efficiency and portability. We demonstrate state-of-the-art accuracy on six publicly available datasets as well as a substantial reduction in processing time, allowing for the analysis of very large images. We also enable InstanSeg to be deployed within popular open-source software tools, ensuring that it can be widely used.

We then extend InstanSeg for the simultaneous segmentation of nuclei and cells in multiplexed fluorescence images. To this end, we develop a neural network architecture for generating three-channel representations of multiplexed images irrespective of the number or ordering of imaged biomarkers. We pair this architecture with InstanSeg and demonstrate state-of-the-art segmentation of multiplexed images on two publicly available datasets.

Furthermore, we demonstrate that InstanSeg can be used as an effective precursor to cell classification by developing an end-to-end cell classification workflow to detect and classify immune cells in kidney biopsies. Finally, we extend our cell classification workflow to multiplexed images allowing for the accurate phenotyping of cells based on biomarker positivity.

Overall, this thesis advances the field of bioimage analysis by providing new computational tools for the segmentation and phenotyping of cells in complex histology images. Our open-source implementations contribute to the improvement and democratisation of state-of-the-art bioimage analysis tools available to biologists.

Lay Summary

Understanding the composition and function of cells in human tissue is crucial for studying a number of diseases, including cancer. Biologists and clinicians now have access to powerful microscopes that can take highly detailed images of the cells that make up tissue. However, these images can be huge and complex, which makes them challenging to interpret. This thesis focuses on developing automated tools for finding and identifying cells in microscopy images. The aim is to help researchers have a better understanding of the cells that make up tissues and how these can change during disease.

First, we introduce InstanSeg, a machine learning algorithm that not only detects cell nuclei but also finds their boundaries. We show that InstanSeg is more accurate than existing algorithms for this task, and is also substantially faster. This means that InstanSeg can be used across huge microscopy images, which can contain over a million cells. We make InstanSeg easily accessible and usable by researchers, including those who do not have any coding experience, through a user-friendly interface.

In a second part, we extend InstanSeg for detecting cells in fluorescence microscopy images. These images can capture many separate signals that reveal the location of important molecules inside cells and tissue. Our eyes can't always perceive all of these signals simultaneously, but this information can be used by computers to assess the identity and function of many types of cells. We developed a new type of deep learning algorithm that can handle images with arbitrarily complex combinations of imaging signals to determine the boundaries of cells. We also extended InstanSeg to detect both nuclei and cell membranes simultaneously for even better separation of cell types.

We then designed a method that can classify types of immune cells involved during tissue inflammation. Some immune cells are so difficult to identify that even experts struggle to tell them apart, if they can do it at all. We managed to train our algorithm using data that was collected automatically, via a chemical reaction, rather than depending on annotations made by expert pathologists. This allowed our method to be highly accurate, and our solution came joint first in an international competition. Finally, we extended our cell classification work to fluorescence microscopy images, which let us classify many more types of cells in tissue.

Acknowledgements

This thesis is the result of my work within the group of Peter Bankhead and the Biomedical AI CDT at the School of Informatics, University of Edinburgh, funded by UKRI. It would not have been possible without the help of many people.

First and foremost, I would like to thank my supervisor, *Peter Bankhead*, for the opportunity to pursue this project, and for providing support and guidance throughout my PhD studies. Thank you, Pete, for sharing your extensive knowledge of bioimage analysis. Your dedication to open-source research and to helping the biomedical community has been inspiring and will continue to shape my approach to research for years to come.

I am also grateful to my secondary supervisor, *Hakan Bilen*, and my external supervisor, *Andrew Filby*, for sharing their expertise and providing advice that greatly strengthened this work. I also thank *Oisin Mac Aodha* for his contribution to the annual review process.

I would like to extend my thanks to the QuPath team: *Alan O'Callaghan*, *Fiona Inglis*, *Leo Leplat*, and *Laura Nicolas Saenz*, for their support, collaboration and friendship throughout my studies. I am also grateful to visiting Postdoctoral researcher *Pau Carrillo Barberà* for our many discussions on segmentation and image analysis.

During my time within the Biomedical AI CDT group, I was surrounded by many amazing PhD students, including *Barry Ryan*, *Sebestyen Kamp*, *Ben Philps*, *Dominic Phillips*, *Charlotte Merzbacher* and *Hans-Christof Gasser*. I'm sure our friendship and collaboration will persist beyond our studies.

Thank you to my brother Antoine for showing me that a PhD should be thoroughly enjoyed, and for reminding me that I could have cycled across much of the Middle East and Africa in the time it took me to write my thesis.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

The work presented in Chapter 3 is undergoing review at Medical Image Analysis, MICCAI. It was deposited as a preprint: Goldsborough et al. (2024). Author contributions are as follows: T. Goldsborough designed the study, gathered all data, conducted experiments, contributed to the software implementation and drafted the manuscript. B. Philps contributed to the analysis. O'Callaghan, F. Inglis, L. Leplat contributed to the software implementation. A. Filby co-supervised the study. H. Bilen contributed to the study design, revised the manuscript and co-supervised the study. P. Bankhead contributed to the study design, contributed to the software implementation, revised the manuscript and supervised the study.

The work presented in Chapter 4 was deposited as a preprint: Goldsborough, O'Callaghan et al. (2024), and will be submitted for review. Author contributions are as follows: T. Goldsborough designed the study, gathered all data, conducted experiments, contributed to the software implementation and drafted the manuscript. O'Callaghan, F. Inglis, L. Leplat contributed to the software implementation. A. Filby co-supervised the study. H. Bilen contributed to the study design, revised the manuscript and co-supervised the study. P. Bankhead contributed to the study design, contributed to the software implementation, revised the manuscript and supervised the study.

The work presented in Chapter 5 and 6 is not yet submitted for peer-review. Author contributions are as follows: T. Goldsborough designed the study, gathered all data, conducted experiments and drafted the manuscript. P. Bankhead contributed to the study design and data annotation, revised the manuscript and supervised the study.

Thibaut Goldsborough

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Microscopy images have long been recognised to contain a tremendous source of information about the morphological and structural characteristics of cells and tissue. Extracting these features enables us to uncover functional and mechanistic insights into a wide range of biological processes and pathologies, which is crucial for the understanding and treatment of disease. Ever since we have been able to capture microscopy images digitally, biological and clinical research has increasingly relied on computational methods to reliably and reproducibly quantify the information held in these images.

Nevertheless, while we know that images of biological and clinical specimens are rich in meaningful information, developing robust computational methods that can extract all of this information across microscopy modalities and experimental conditions remains a major challenge. The difficulty stems partly from the complexity and subtlety of the biological phenomena that we can observe, but also from the scale of imaging data that we have recently been able to produce. This means that we need algorithms that are not only more sensitive, but also more scalable.

In the last decade, machine learning methods have shown the most promise for meeting these demands. In particular, deep learning has demonstrated that subtle morphological patterns in imaging data can be captured, sometimes surpassing the sensitivity of the human eye (Hekler et al., 2019), (Cifci, Foersch, & Kather, 2022), and some methods are able to do so at scale. However, the widespread advantages that deep-learning methods could bring to bioimaging research are hindered by our ability to integrate these approaches into practical research workflows. Often, deep-learning tools have been developed to improve accuracy in highly-specific use cases at the cost of broader usability and real-world applicability.

There remains a pressing need for accurate and scalable automated analysis methods that are designed and tailored for the practical needs of biologists and clinicians. Such solutions must not only maintain high performance but also be straightforward to implement, so that recent computational advances can directly support real-world biological and clinical research.

This thesis tackles the problem from the ground up, by focusing on the fundamental unit of life: the cell. Cells are central to biological research and are crucial for understanding both normal physiology and disease. In particular, the detection of cells and their boundaries is a key step in many bioimage analysis pipelines. Cell detection methods need to be detailed enough for the extraction of features, such as morphology or intensity measurements, which are often indicative of cell type, function and physiology (Chandrasekaran et al., 2024). Capturing the boundaries of cells is also necessary to study intracellular composition and processes (Cho et al., 2022).

Furthermore, cell detection methods need to be scalable enough to capture and query thousands or millions of cells, which is necessary to identify very rare cell types (Ndacayisaba et al., 2022), rare cellular events (Aubreville et al., 2023), to study cell interactions (Brbić et al., 2022) and understand complex tissue architectures such as tumours (Jackson et al., 2020). Accurate and efficient detection algorithms can generalise to other structures, from nano-scale subcellular spots (Dominguez Mantes et al., 2025) to millimetre-scale tissue regions (Sirinukunwattana et al., 2017), underpinning a whole range of downstream analysis workflows and applications.

This thesis sets out to advance biological and biomedical research by developing novel methods to segment and phenotype cells. We aim not only to equal or surpass the state-of-the-art in terms of accuracy across multiple imaging modalities using common benchmarks, but also to substantially improve efficiency and usability. Both aspects are crucial to meet the real-world needs of researchers, and maximize the information we can gain from bioimaging datasets.

1.2 Research outputs

1.2.1 Preprints and published work

- **Goldsborough, T.** et al. (2024) 'InstanSeg: an embedding-based instance segmentation algorithm optimized for accurate, efficient and portable cell segmentation'. arXiv. Available at: https://doi.org/10.48550/arXiv.2408.15954.
  [In preparation for journal submission]
- **Goldsborough, T.** et al. (2024) 'A novel channel invariant architecture for the segmentation of cells and nuclei in multiplexed images using InstanSeg'. bioRxiv, p. 2024.09.04.611150. Available at: https://doi.org/10.1101/2024.09.04.611150.
  [In preparation for journal submission]
- Hunter, B, Nicorescu, I, Foster, E, McDonald, D, Hulme, G, Fuller, A, Thomson, F, **Goldsborough, T.**, et al. (2023) 'OPTIMAL: An OPTimized Imaging Mass cytometry AnaLysis framework for benchmarking segmentation and data exploration'. Cytometry. https://doi.org/10.1002/cyto.a.24803
  [Contributions: Methodology, Benchmarks]
- Co-authorship on the upcoming *Machine-learning for Optimal detection of iNflammatory cells in the KidnEY (MONKEY) | MIDL 2025*

1.2.2 Other outputs

Workshop presentations

- 'Deep-learning methods of cell segmentation in microscopy images' | QuPath Workshop: From Samples to Knowledge | October 2023 | La Jolla Institute of Immunology, San Diego, US
- 'Introducing InstanSeg, a deep-learning-based method for the segmentation of cells in brightfield and fluorescence images' | October 2024 | Virtual I2K
- 'InstanSeg, un algorithme de segmentation cellulaire pour les images en brightfield et en fluorescence' | November 2024 | Institut Cochin, Paris

Conference presentations [selected]

- 'InstanSeg, a novel cell segmentation algorithm for analysis of brightfield and multiplexed images' | November 2024 | Frontiers in Bioimaging, Royal Microscopical Society, Oxford, UK
- 'Multiplexed Image Analysis' | Oct 2023 | Quantitative Bioimaging Society San Diego, US
- 'Deep Morphological Analysis of Single-Cell Images Captured by Imaging Flow Cytometry' | November 2022 | Crick Bioimage Symposium, London, UK

Poster prizes

- *Best poster Awards by popular and jury vote* | I2K 2024 From Images to Knowledge | October 2024
- *Best poster Award by popular vote* | Annual UKRI CDT Poster Showcase | May 2025

Public competitions

Machine-learning for Optimal detection of iNflammatory cells in the KidnEY (MONKEY) | MIDL 2025

- 1st place (leaderboard 1) and 1st place (leaderboard 2) in the public test set
- 2nd place (leaderboard 1) and 1st place (leaderboard 2) in the private test set

Pypi Packages

instanseg-torch (Version 0.0.9) [Software]. (2025, April 17). PyPI.
https://pypi.org/project/instanseg-torch/

1.3 Thesis structure and contributions

The contributions of this thesis can be summarised as follows:

Chapter 3: InstanSeg: an embedding-based instance segmentation algorithm optimised for accurate, efficient and portable cell segmentation

- A novel embedding-based method, InstanSeg, for the instance segmentation of nuclei in brightfield microscopy images.
- A demonstration that InstanSeg provides state-of-the-art performance for nucleus segmentation in brightfield images.
- A Python library for the scalable segmentation of cells in microscopy images using InstanSeg, with support for images that are larger than RAM.
- A self-contained implementation of InstanSeg that can be run in open-source software QuPath and Fiji.

Chapter 4: A novel channel invariant architecture for the joint segmentation of cells and nuclei in multiplexed images using InstanSeg

- A novel channel-invariant architecture capable of generating fixed representations of highly multiplexed fluorescence images irrespective of the order and number of imaged biomarkers.
- An extension of InstanSeg for the joint segmentation of nuclei and cells in fluorescence images.
- A demonstration that InstanSeg + ChannelNet provides state-of-the-art performance for nucleus and cell segmentation in multiplexed images.
- A proof-of-concept showing that a SAM-based encoder can be used within InstanSeg for generalization across imaging modalities, in both tissue and cell cultures.

Chapter 5: The MONKEY challenge: cell classification in brightfield images using weakly-annotated labels

- The development of a weakly-supervised phenotyping pipeline for the detection of immune cell types in brightfield images of renal biopsies.
- A demonstration of state-of-the-art performance, ranking joint-first on the MONKEY Challenge leaderboard, showing that segmentation and classification can be modularised effectively.

Chapter 6: Marker-agnostic cell phenotyping in multiplexed fluorescence images

- An investigation into the impact of cell segmentation methods on downstream feature representation and unsupervised clustering.
- A synthetic data generation pipeline for simulating cellular biomarker expression.
- A new architecture for the population-aware prediction of biomarker positivity in cells.

Chapter 2

Background

2.1 Microscopy for cell and tissue phenotyping

Microscopes are central to biological and biomedical research, enabling the direct visualisation of tissues, cells, and subcellular structures. Microscopy enables cell and tissue phenotyping, whereby morphological characteristics are assessed to distinguish between cell types and functions, identify pathologies and discover underlying biological mechanisms. In this section, we review the historical development of microscopy, outline common imaging modalities used for tissue and cell phenotyping, and discuss current challenges that motivate the development of novel computational approaches to the analysis of microscopy images.

2.1.1 A brief history of microscopy

For millennia, our understanding of the natural and physical world was fundamentally restricted by the limitations of our eyes. Although simple lenses may have been used during antiquity, it was not until the 16th century that convex glasses and eventually compound microscopes were first used for the study of natural phenomena (Singer, 1914). With the gradual technological improvement in the manufacturing of glass lenses, early microscopy pioneers such as Robert Hooke and Antonie van Leeuwenhoek began to uncover a new microscopic world of *animalcules*: protozoa, cells, bacteria and even organelles. Hooke's *Micrographia* and Leeuwenhoek's meticulous description of his discoveries in letters to the Royal Society of London drew considerable interest from the scientific community (Singer, 1914), which persists to this day.

Over the next centuries, microscopy saw gradual but significant improvements in magnification power, lighting and sample preparation. One notable development was the discovery of chemical dyes that could reveal the chemical and molecular composition of biological samples: Prussian blue could detect the presence of iron, Gram staining enabled the identification of bacteria and haematoxylin could reveal nucleic acids. At the start of the 20th century, Paul Ehrlich found that some chemical stains were specific enough to segregate immune cell types, or selectively stain infectious bacterial cells while ignoring host cells, a discovery which was rewarded with a Nobel Prize in 1908 (I. H. Hussein & Raad, 2015). The combined use of multiple stains allowed histologists and pathologists to discern the cellular organisation of tissues, classify cell types and understand disease. Many of the stains developed in the 19th and 20th century remain in use to this day.

Chemical stains show impressive specificity, yet they are insufficient to reveal the huge complexity and diversity in the molecular composition of tissues. The next cornerstone in the exploration of the microscopic world was the development of immunolabeling using fluorescent antibodies, attributed to Albert Coons in 1941 (van Ooij, 2009), and later adapted to standard light microscopes (Avrameas & Uriel, 1966). These enabled the visualisation of virtually any protein and its localisation in biological tissues. Antibody-based stains could be multiplexed to simultaneously observe a number of biological targets, improving our ability to characterise cell types and function, understand biological processes and uncover mechanisms of disease.

In parallel with improvements in sample preparation, fixation and labelling, light microscopes went through a number of revolutions throughout the 20th century. While standard brightfield microscopes used the attenuation of light through a sample to create contrast, darkfield microscopes used the scattering of light and phase-contrast microscopes used interference patterns arising from differences in refractive index to visualise otherwise transparent samples (Amos, 2000). Fluorescence microscopy, which became increasingly powerful throughout the latter half of the century, allowed for even greater specificity and contrast by exploiting the emission of light from fluorophores. It was eventually demonstrated that light microscopy could resolve structures smaller than the wavelength of visible light, using super-resolution techniques such as PALM and STORM (Henriques, Griffiths, Hesper Rego, & Mhlanga, 2011) , further expanding the range of biological mechanisms we could observe and characterise (Danial, 2025).

In the mid 20th century, optical images of samples mounted on microscope slides were first digitised. The digitisation process samples the optical image to produce an integer array of sample points, where each value corresponds to the amount of detected light at a given point. These digital representations enabled display on computer screens and downstream quantitative analysis (Merchant & Castleman, 2022). Initially, microscopy images were restricted to small field of views of perhaps 1024 x 1024 pixels. Multiple fields of view could be imaged at different focal depths or stitched to produce larger composite images. The introduction of slide scanners in the 1990s allowed entire microscope slides to be digitised and stored as Whole Slide Images (WSIs), which meant that samples could be visualised, shared and analysed virtually after acquisition, often referred to as virtual microscopy (Parwani, 2022). A single WSI capturing an entire microscope slide could contain 100k x 200k pixels, equivalent to around 50 GB. Today virtual microscopy has given rise to digital pathology, in which research and even diagnostic workflows increasingly rely on the analysis of digitised WSIs rather than by direct examination under a microscope (Pantanowitz et al., 2018). Digitisation and virtual microscopy not only improved reproducibility, accessibility and collaborations between institutions, it allowed for the development of computational methods for the quantitative analysis of biological and biomedical samples.

2.1.2 Imaging modalities for cell and tissue phenotyping

Modern biological and clinical research relies on a range of microscopy modalities, each with unique characteristics. We briefly describe some of the main modalities and their use for biological phenotyping, and illustrate some in examples in Fig. 2.1.

Brightfield microscopy requires external lighting and uses light attenuation to generate contrast (Bancroft, 2019). Brightfield images are characterised by a white background and often require stains for the visualisation of biological structures. The haematoxylin and eosin (H&E) stain, in use for over a century, is an unspecific staining method for exposing nuclei and cells. Immunohistochemistry (IHC) can reveal specific proteins in brightfield images, using chromogens such as DAB, which produces a brown coloured precipitate at the site where the target protein is present.

Fluorescence microscopy depends on fluorophores to emit light. Fluorophores can be used for the visualisation of cellular structures and molecules with high specificity. A commonly used fluorophore is DAPI, which reveals DNA and is routinely used to label cell nuclei in fixed samples. Fluorescence microscopy supports complex techniques such as three-dimensional and super-resolution imaging methods.

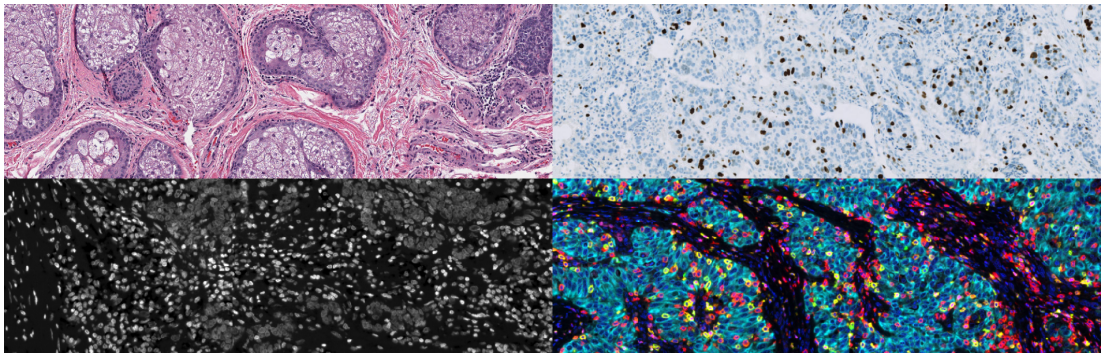

Figure 2.1: Common microscopy modalities used in biological research, shown clockwise: (top-left) Brightfield image of haematoxylin and eosin (H&E) stained tissue revealing nuclei and tissue structures (CMU-2.svs, OpenSlide); (top-right) Brightfield immunohistochemistry (IHC) image showing DAB (brown) stain revealing proliferating cells (OS-2.ndpi, OpenSlide); (bottom-right) Fluorescence microscopy image of nuclei stained with DAPI; (bottom-left) Multiplexed fluorescence image visualising multiple biological targets, such as cell type markers, rendered in RGB for display. Fluorescence images from Aleynick et al. (2023)

Multiplexed fluorescence imaging allows for the simultaneous visualisation of a number of fluorophores through techniques such as cyclical fluorophore excitation, cyclical staining and imaging. Digitised multiplexed images are atypical as they often contain a higher number of channels than typical 8-bit RGB images.

Phase-contrast microscopy reveals contrast in transparent, unstained samples by converting phase shifts in transmitted light into changes in intensity. This technique is useful for observing living cells in culture, as it avoids the need for staining or fixation. Digitisation produces one channel (greyscale) images.

Digitised microscopy images contain a wealth of information capturing the cellular composition of tissues which can elucidate biological pathways and help uncover disease mechanisms. However, digitised images pose unique challenges for computational processing and analysis, as we explore in the next section.

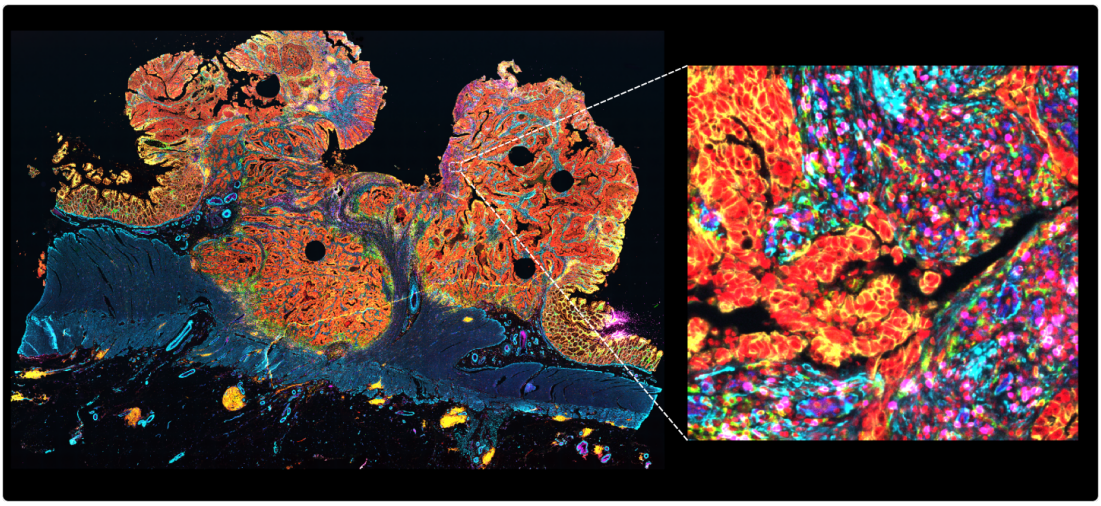

Figure 2.2: Example of a multiplexed whole-slide image (WSI). The left panel shows a low-resolution overview of the full WSI (dimensions: 19 channels, 49,495 × 71,344 pixels) with an uncompressed size of 140 GB. A small crop is shown at higher resolution, capturing nuclei (in red) along with several other cellular markers. The full WSI contains over one million cells, demonstrating the need for scalable bioimage analysis algorithms. Image data from J. Lin et al. (2023)

2.2 Fundamentals of bioimage analysis

2.2.1 Challenges and aims in bioimage analysis

Digital images are composed of pixels, or picture elements, each representing a scalar value typically associated with light intensity. These pixels are frequently organised in numerical arrays that maintain the spatial arrangement of the input signal. Intrinsically, even small images have exceptionally large dimensionality, requiring efficient methods that can summarise information. Modern multiplexed fluorescence WSIs are a good example of this challenge, as a single image can contain over a hundred gigabytes of data capturing millions of cells across dozens of channels (Fig. 2.2). To capture the complexity and subtlety of biological mechanisms captured in such images, analysis methods need to be both extremely efficient and highly accurate.

The study of digital images relies on computational methods for *image processing*, where pixel values are manipulated to aid downstream interpretation, and *image analysis*, where measurements are extracted to summarise the information content of an input signal (Bankhead, 2022). Processing and analysing images often relies on simple arithmetic or algorithmic manipulations that can be applied sequentially.

The output of an image analysis pipeline can vary considerably depending upon the application: for example, a single scalar (e.g. the number of cancer cells in an image), a vector (e.g. morphological features of a cell) or another image (e.g. a heatmap of cancer cell density in a biopsy).

2.2.2 Principal objectives of bioimage analysis

Bioimage analysis can be extremely varied, due to both the diversity of imaging modalities and the complexity of biological research objectives. Despite the complexity, most analysis pipelines share common strategies and two steps in particular are almost ubiquitous: (1) the detection of areas or objects, and (2) the extraction of measurements from these detected objects for downstream analysis.

The first step is termed segmentation, and involves partitioning an image into objects or regions of interest. This can take the form of semantic segmentation, where every pixel is assigned to a category (e.g., foreground, tumour, nucleus). This may determine that a pixel is part of a cell, for example, but not which specific cell. Instance segmentation goes a step further by distinguishing individual objects, which may be represented using contours or pixel labels. In bioimage microscopy analysis, instance segmentation often concerns nuclei and cells but can also include larger structures such as blood vessels and crypts: essentially any entity for which a boundary can be defined.

Typically, object segmentation is followed by the extraction of measurements and classifications for further analysis. Object measurements can be morphological (e.g. area, circularity) or intensity-based (e.g. mean, maximum, standard deviation of pixel values). These are often used for object classification. This can be as simple as applying a threshold to a single measurement (e.g. to identify cells positive for a particular marker based upon staining intensity only) or involve training a machine learning classifier on multiple extracted features (e.g. using random forests to distinguish tumour from non-tumour cells). Multiple classification steps may be required. Finally, object measurements and classifications underpin downstream analysis and querying which can be simply object counting (e.g. percentage of positive tumour cells) or can be substantially more complex (e.g. using spatial statistics).

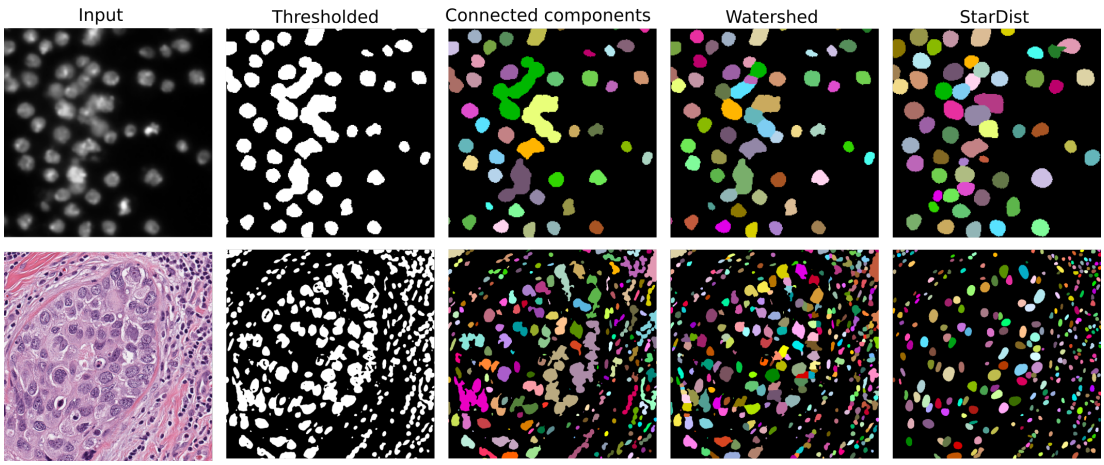

Figure 2.3: Conventional image analysis pipelines such as thresholding, connected component analysis and watershed transformations can struggle to separate touching nuclei in microscopy images when compared to deep-learning based methods such as Stardist (Schmidt et al., 2018). Top image from Caicedo et al. (2019), bottom image from Kumar et al. (2017).

2.2.3 Conventional methods for image analysis

In a conventional pipeline, the workhorses for image segmentation are filters and thresholds. A filter computes a new value for each pixel based on a weighted sum of the values in its local neighbourhood. The weights in this operation are known as a kernel, and the process of applying it across the spatial dimensions of an image is called a convolution. Some kernels can perform operations like blurring to reduce noise or edge detection to identify object boundaries. Following a single or multiple filtering operations, a threshold can be applied to create a binary semantic segmentation map. This map can be split into distinct objects using a connected-component algorithm if objects are already separated or a watershed transform if the objects are touching, as illustrated in Fig. 2.3.

However, developing these conventional pipelines has significant drawbacks. The process demands substantial domain knowledge and manual effort to design and tune the sequence of filters and algorithms. The resulting workflows are often brittle and fail to generalise across different image modalities, biological samples, or experimental conditions. For many complex bioimage applications, conventional methods are simply insufficient, producing unreliable results as we illustrate in Fig. 2.3. This created a clear need for methods that were both more powerful and robust, but could automatically learn optimal processing operations directly from data, which led to the widespread adoption of machine learning.

2.3 Deep-learning for bioimage analysis

While conventional methods can excel at many tasks in bioimage analysis, they often require manual tuning of parameters and heuristics. In practice, conventional methods can fail to generalise across experimental conditions and imaging modalities. Of concern, the results from conventional image analysis pipelines are vulnerable to interpreter bias, such as the setting of thresholds. These limitations have motivated a shift towards machine learning solutions, in which an algorithm learns a mapping between an input and a desired output from a set of annotated examples.

2.3.1 Optimisation and neural network training

Canonically, machine learning solves the problem of approximating a mapping $f : X \mapsto Y$ given N paired $(x_i, y_i)_i^N$ training samples from $X^{\text{train}} \subset X$ and $Y^{\text{train}} \subset Y$ by a parametrised function f_θ. During the training process the parameters θ of the model are adjusted to minimise a loss function $\mathcal{L}(f_\theta(x_i), y_i)$, measuring the discrepancy between the model outputs and the ground truth (Goodfellow, Bengio, Courville, & Bengio, 2016).

The most widely implemented algorithm to update the parameters θ so as to minimise the training loss $\mathcal{L}(f_\theta(x_i), y_i)$ is stochastic gradient descent (SGD). Under this approach, the parameters of a model are updated in the opposite direction of the gradients ∇_θ over a randomly sampled batch of N_b training inputs

$$\theta_{n+1} = \theta_n - \alpha \frac{1}{N_b} \sum_{i=1}^{N_b} \nabla_\theta \mathcal{L}(f_\theta(x_i), y_i) \tag{2.1}$$

with a small learning rate α (e.g. α = 0.001). This optimisation routine requires both the loss function L and the model g_θ to be differentiable. When the parameters of the model are applied in sequential layers, such as in feed forward neural networks, the gradients ∇_θ can be calculated with respect to each layer by applying the chain rule successively backwards through the model, called back-propagation. In practice, algorithms like Autograd (Baydin, Pearlmutter, Radul, & Siskind, 2018) can track and compute gradients across tensor operations, so the gradient of any quantity with respect to any other can be automatically computed without having to manually imple-

ment the chain rule. Many variants of SGD have been suggested, a popular example is Adam (Kingma & Ba, 2017), which normalises gradients throughout the model by tracking running estimates of the mean and variance of each gradient, encouraging homogeneous training speeds in different parts of a model.

It is typical to train several different models on a training set and select the best performing one based on performance on a $X^{\text{validation}}$ set using a performance metric P (Goodfellow et al., 2016). Finally a model can be evaluated on final held-out set X^{test}, to obtain an estimate of its real-world performance.

2.3.2 Machine learning models and layers

The linear layer

Perhaps the most important module in machine learning is the linear layer, which benefits from decades of research and hardware optimisations. The layer applies an affine transformation to an input vector. The fully connected layer implements an affine transformation of an arbitrarily shaped tensor into an arbitrarily shaped output tensor. In practice, fully connected layers can learn geometric transformations such as projections, rotations and translations. Fully connected linear networks can be assembled by applying fully connected layers sequentially, however, models that only contain sequential affine operations can only output affine transformations of an input. For many non-linear tasks, it is common to introduce non-linear operations between layers, a popular example is the Rectified Linear Unit (ReLU) (Glorot, Bordes, & Bengio, 2011).

While linear models can take as input an arbitrarily shaped tensor, they scale very poorly with input size. For example, the smallest fully connected model that takes as input as a small RGB image of 500 by 500 pixels and outputs an image of the same size, would require at least $5.6e^{11}$ parameters. Apart from computational complexity, this model would ignore the highly structured local spatial information of most images.

Convolutional neural networks

Since their introduction in the 1980s, convolutional neural networks (CNNs) have become foundational to the field of computer vision. CNNs are typically composed of convolutional, pooling and upsampling operations that are performed both in parallel and sequentially, these operations are often combined in blocks and the ordering of these blocks is termed an architecture. By construction, the operations applied to input pixels are identical irrespective of the pixel coordinates. This leads to an interesting bias termed *translational equivariance*. While CNNs were first developed for natural images (LeCun, Bottou, Bengio, & Haffner, 2002), (Krizhevsky, Sutskever, & Hinton, 2012) inductive bias makes CNNs especially well suited for microscopy images.

Convolutional layers, inspired by conventional kernel operations, apply a set of trainable weights across the spatial dimensions of an image. In its simplest form, the convolutional layer is analogous to applying a fully connected layer with k^2 input features and f output features to each $k \times k$ pixel neighbourhood, producing f feature maps of the same spatial dimensions as the input image. Similar to manually crafted kernels, this operation can learn to identify edges at object boundaries, reduce noise, or detect local spatial patterns. Besides k and f, each convolutional layer depends on a set of hyperparameters such as convolutional stride and kernel dilation. These parameters determine the size of the pixel neighbourhood affecting each pixel and can either increase or reduce the spatial dimensions of the resulting feature maps.

The pooling layer is another way to reduce the spatial dimensions of an image. The layer aggregates information across small pixel neighbourhoods, commonly by computing local maxima. For example, a 2×2 max pooling layer would replace 4 pixels with the maximum value, decreasing both the width and height by a factor of 2. This operation can be likened to an activation layer, as it can suppress the contribution of less activated features.

The upsampling layer increases the spatial dimension of a feature map. It can either be trainable using transposed convolutions or non-trainable using interpolation. Uncommon in most machine learning architectures for increasing the sparsity of feature maps, this layer is necessary to either recover the spatial dimensions of an input image from lower resolution feature maps or for super resolution, where the goal is to reconstruct a high-resolution image from a low-resolution input.

Skip connections allow features from earlier layers to directly attend downstream layers. This allows spatially detailed information, which can be lost through pooling or repeated convolutions, to be permeate through a network. Skip connections are usually implemented by concatenation or by summation.

Fully convolutional networks (Long, Shelhamer, & Darrell, 2015) are networks that are composed only of convolutional, pooling or upsampling layers, without relying on fully connected layers. By construction, convolutional networks apply the same filters and operations across the spatial dimensions, which confers several advantageous properties. First, these networks are not constrained to a fixed input size, allowing a model to be trained on conservatively sized images in limited computational settings or with a higher batch size, but can be used downstream on larger images without affecting accuracy. Second, this property can be used to stitch overlapping tiles of very large images while minimising tiling artefacts - which is an important consideration as bioimages increase in size. Finally, translational equivariance is especially well suited to the field of digital microscopy as images often don't have meaningful ups and downs, left or rights, in contrast to many natural images. This alleviates a model from having to learn to recognise similar features on opposite sides of an image independently multiple times, reducing the number of weights and risk of over-fitting on often limiting biomedical data.

Key Work: The U-Net architecture

It's hard to overstate the importance of the U-Net architecture (Ronneberger, Fischer, & Brox, 2015) in the field of computer vision, especially for the task of image segmentation in the biomedical world. U-Net is a fully convolutional network and comprises of two parts: (1) an encoder, which sequentially halves the spatial dimensions of the input image while doubling the number of features, and (2) a decoder, which sequentially doubles the spatial dimension and halves the number of feature maps. Crucially, skip connections between the encoding blocks and the decoding blocks allow fine-scale structures to permeate through the network. It is thought that lower resolution feature maps capture more high-level properties of an input image, which are allowed to propagate through the decoder layers. Owing to its conceptually simple symmetric architecture, translational equivariance and high performance, the U-Net architecture and its variants such as Attention-UNet (Oktay et al., 2018) and UNet++

(Zhou, Siddiquee, Tajbakhsh, & Liang, 2019) have become a common baseline in many biomedical applications, with many recent competition winners often using a variant of the architecture (Ulman et al., 2017), (Isensee, Jaeger, Kohl, Petersen, & Maier-Hein, 2021), (Stringer & Pachitariu, 2024) in biomedical settings.

However, in some computer vision applications, especially natural images in which the interaction of objects within an image can be extremely complex, U-Nets have shown some limitations in their ability to capture global context and model long-range spatial dependencies.

Vision transformers

Even a short review of deep-learning architectures in computer vision would not be complete without recognising the importance of the transformer architecture (Vaswani et al., 2017). This architecture was initially developed for text processing models and departs entirely from the convolutional approach, but gained popularity in computer vision after the seminal work of (Dosovitskiy et al., 2021). In short, an image of pre-defined spatial dimensions is split in $k \times k$ pixel patches, flattened, and projected into an embedding space using a linear layer. At this stage a learned positional embedding is added to the patch embeddings, which serve as input to a transformer encoder. This transformer model contains alternating multiheaded self-attention blocks and fully connected layers, allowing patch embeddings to attend themselves and others. These blocks excel at capturing complex interactions between image elements that are far apart in the image plane.

Of note, the 1-D spatial embeddings are initialised randomly, meaning that the spatial relations between patches have to be learnt by the model, in contrast to CNNs. An advantage of vision transformers over their convolutional counterparts is the larger model capacity which scales effectively with increasing compute and training images. ViTs are also understood to capture high-level interactions between patches better than CNNs.

In the biomedical setting, ViTs and U-Nets have mostly played a complementary role. While U-Nets have proved easy to train from scratch for a diverse range of biomedical applications, ViTs tend to require more training annotations and more advanced hardware, but their ability to learn from large, diverse datasets has made them central to many modern, general-purpose vision models.

2.3.3 Loss functions and performance functions

Loss functions for semantic segmentation

The choice of loss function to be minimised plays a critical role during neural network training, and has been accompanied by a large body of research (Ma et al., 2021). For semantic segmentation, which is a pixel-wise classification problem, loss functions can be broadly categorised into two main families: the distribution based loss functions, such as standard cross-entropy (CE), and the geometry based loss functions, such as Dice loss (Milletari, Navab, & Ahmadi, 2016). While there seems to be no clear consensus in the literature as to which type of loss function is better (B. Liu, Dolz, Galdran, Kobbi, & Ayed, 2024), it is agreed that each loss function is associated with specific biases.

Standard cross-entropy (CE) loss matches the predicted probability distribution to the ground-truth labels on a per-pixel basis. In the binary case, CE is defined as:

$$\mathcal{L}_{\mathsf{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right],$$

where $\hat{y}_i$ denotes the predicted probability, y_i the binary ground truth label, and N the total number of pixels. Classes or regions occupying smaller pixel areas contribute less to the overall loss, potentially leading to under-segmentation of small structures. To address this, weighted cross-entropy (WCE) and focal loss (T. Y. Lin, Goyal, Girshick, He, & Dollár, 2017) are commonly used to mitigate class imbalance and have become popular in natural image segmentation tasks.

In contrast, Dice loss maximises the relative overlap between predicted and ground-truth regions. It is a differentiable reformulation of the Dice Similarity Score (DSC). In the binary case, the Dice loss proposed by (Milletari et al., 2016) is written as:

$$\mathcal{L}_{\mathsf{Dice}} = 1 - \frac{2\sum_{i=1}^{N}\hat{y}_i y_i}{\sum_{i=1}^{N} y_i^2 + \sum_{i=1}^{N}\hat{y}_i^2}.$$

Dice loss operates on the overlap between predicted and true regions, reducing the influence of large background regions and helping to preserve small structures during training. Equivalent loss functions, such as the Lovász-hinge loss (Berman, Triki, & Blaschko, 2018), have also been proposed to better handle imbalanced regions by directly optimizing the Intersection-over-Union (IoU) instead of DSC. These losses are more commonly used in medical and biological images, where pixel-wise class imbalances can be extreme.

Unfortunately, pixel-wise semantic segmentation losses rely on fixed label targets. They fail in instance segmentation tasks, when the number of target objects varies from image to image, and the ordering of object labels is arbitrary. This makes instance segmentation a substantially more challenging task than semantic segmentation. As we explore in the next section, instance segmentation methods have had to use other types of loss functions, such as regression or embedding-based losses that implicitly segregate pixels into separate instances.

Performance metrics assess segmentation performance

Ultimately the choice of loss function and associated biases should depend on the needs of downstream applications and be captured by an appropriate performance metric. While loss functions and metrics are highly related, they play different roles: losses need to be differentiable and prioritise smoothness, while metrics are usually chosen for the interpretable assessment of final performance, they are often non-differentiable and computed after thresholding. Common semantic segmentation metrics are IoU (also called Jaccard index) and DSC, which measure the pixel overlap between a predicted and ground truth region. For instance segmentation, metrics often involve Average Precision (AP) and average F1 scores over all the objects in an image. These take the ratio of true positives (TP), false positives (FP) and false negatives (FN) detections over a range of IoU matching thresholds (e.g. 0.5 to 0.9). Precision and F1 score are defined by

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{2.2}$$

$$F_1 = \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FP} + \text{FN}}. \tag{2.3}$$

While AP seems to be the conventional metric in the natural image community, F1 score is by far the most common metric in medical imaging. Similarly, the medical imaging community often replaces IoU with the related DSC (Maier-Hein et al., 2018). The reason for the discrepancy seems to stem from simple community preferences (Maier-Hein et al., 2024).

It is important to recognise the limitations of common detection metrics. First, they typically ignore model confidence, treating predictions as binary decisions rather than probabilistic estimates. Secondly, they are undefined when there are no ground truth objects in an image (e.g. background region of a microscopy slide). Third, they place a strong emphasis on detection/localisation rather than on the fidelity of shape or boundary. As a result, subtle morphological features (e.g., cellular protrusions) may be overlooked by these metrics. On the other hand, alternative boundary-based metrics, while more sensitive to shape detail, can be highly affected by small annotation inconsistencies (Maier-Hein et al., 2024).

2.4 Objective functions for instance segmentation

This section outlines and evaluates common instance segmentation methods found in literature. We treat cell and nucleus segmentation and segmentation of natural images largely interchangeably, as the two tasks are fundamentally similar: both require assigning pixels to individual object instances. Substantial contributions have come from both traditional computer vision and the bioimage analysis community, and methods developed in one domain may translate readily to the other. Instance segmentation in bioimage analysis typically need to scale to high object densities over very large images (such as whole-slide images), requiring efficient handling of crowded, touching, or overlapping instances. Unlike natural image datasets, bioimage datasets usually involve only a few semantic classes (e.g., cells, nuclei, tissue regions), though these may exhibit considerable morphological variability..

2.4.1 Proposal-free methods

Instance segmentation aims to assign pixels to individual instances. A common target for this task is a labelled map $\boldsymbol{L}$ where background pixels have the value 0 and foreground pixels are labelled $1, 2..., K$ for each of the K individual instances in the image. While this may superficially resemble a classification task, the ordering of the label values is arbitrary and cannot be used as a direct target during training. The field of instance segmentation has often focused on finding surrogate training targets $\tilde{\boldsymbol{L}}$ that are equivalent to a labelled map.

As an educational example, we consider the simplest surrogate objective: a binary target $\tilde{\boldsymbol{L}}_{\text{binary}}$ where foreground pixels are labelled 1, while the background and boundaries between instances are labelled 0.[1] In this formulation, the task becomes a classification task and standard loss functions such as standard CE can be used. Once the parameters of a model have been optimised to minimise the loss function, the model outputs $f_\theta(\boldsymbol{x})$ should resemble $\tilde{\boldsymbol{L}}_{\text{binary}}$. A labelled output $\hat{\boldsymbol{L}}$ can be recovered by non-differentiable steps such as thresholding followed by a connected-component algorithm. Finally, an instance segmentation metric can be used to compare $\hat{\boldsymbol{L}}$ and the ground truth $\boldsymbol{L}$.

For illustrative purposes, we train and evaluate a model on a single image using the above paradigm and show the results in Fig. 2.4. After training, the model outputs closely resemble the binary training target. Yet, the recovered $\hat{\boldsymbol{L}}$ differs substantially to the ground truth $\boldsymbol{L}$, due to the poor separation of instances. This example serves as an illustration that finding a differentiable loss function in which $\mathcal{L}(f_\theta(x_i), y_i) \rightarrow 0 \implies P(\hat{\boldsymbol{L}}, \boldsymbol{L}) \rightarrow 1$ is not sufficient for finding an effective instance segmentation method.

Over the last decade, computer vision researchers have developed a wide range of instance segmentation paradigms to ultimately seek a better correspondence between the training objective and the performance metric. One branch of solutions, often referred to as the proposal-free methods, have sought to modify the above paradigm to optimise the separation of instances. A number of diverse and sometimes highly creative proposal-free methods have been suggested which we will briefly cover in the following section. To compare the methods, a number of important properties have to be considered:

---

1. This conversion can be obtained by label-wise erosion followed by thresholding above 0. The conversion is *usually* reversible using a connected-component algorithm followed by dilation, but may cause thin or small structures to disappear

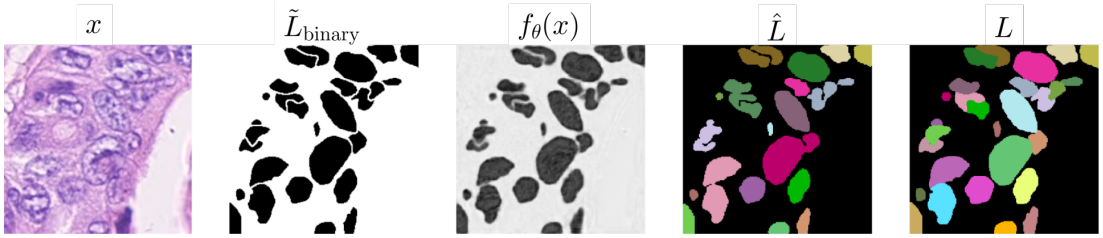

Figure 2.4: An educational example illustrating the difficulty of aligning training objectives for instance segmentation. From left to right, an input image $\boldsymbol{x}$, a binary training target separating instances $\tilde{\boldsymbol{L}}_{\text{binary}}$, a model's prediction $f_\theta(\boldsymbol{x})$, the recovered instance segmentation map $\hat{\boldsymbol{L}}$ and the ground truth instance segmentation map $\boldsymbol{L}$. Despite the model's ability to predict the binary target with high accuracy, the recovered segmentation map differs substantially from the instance segmentation target. Input image from Graham et al. (2019).

1. **Accuracy:** How well the training objective aligns with the performance metric.
2. **Robustness:** The sensitivity of the training objective to labelling mistakes, including variation in the location of the instance boundaries.
3. **Postprocessing speed:** The computational complexity of processing the model outputs into a labelled map.
4. **Hyperparameters:** The number and sensitivity of parameters in the postprocessing steps that cannot be optimised by gradient descent.
5. **Portability:** How easily the postprocessing steps can be deployed for use by an end user.

The now famous U-Net (Ronneberger et al., 2015) paper used a similar training objective to the one described above. The key modification was to introduce a pixel-wise weighting to the cross entropy loss function to penalise model predictions at the instance boundaries, encouraging the separation of objects. This formulation allowed the method to win the ISBI cell tracking challenge (Solórzano, Kozubek, Meijering, & Barrutia, 2015) by a large margin. A year later DCAN (Chen, Qi, Yu, & Heng, 2016) instead predicted the label boundaries in a separate output channel to facilitate the separation of touching objects in postprocessing. Guerrero-Pena et al. (2018) introduce a more complex weighted loss regime to better encourage instance separation. The main limitation of these methods is their high sensitivity to ambiguity, mistakes or inter-labeller variability in the ground truth annotations. These issues are ubiquitous

in public microscopy datasets due to blurry image acquisition and the thickness of samples. As a consequence, the field sought more robust postprocessing functions that were resilient to single pixel mistakes at object boundaries, and training objectives that could optimise them.

A key improvement was the transition from a binary prediction target to the regression of a distance map $\tilde{\boldsymbol{L}}_{\text{distance}}$, in which each foreground pixel is mapped to a distance to the closest instance boundary. This was implemented by Naylor, Laé, Reyal, and Walter (2018) and in Mesmer (Greenwald et al., 2022), still widely used for the segmentation of cells and nuclei in fluorescence images. Crucially, this formulation was much less sensitive to single pixel variation in the object boundaries. For convex objects, this objective tends to favour a single local maximum at the centre of each object. During postprocessing, labelled instances can be recovered using a watershed algorithm, rather than a connected component algorithm, which is more robust to incorrect model predictions at object boundaries. The drawback is that the watershed algorithm depends on finding a single seed for each instance. The formulation suggested by Naylor et al. (2018) and Greenwald et al. (2022) would produce multiple local maxima for non-convex objects. One recent attempt to solve this was proposed by Z. Lin et al. (2023) which involved the regression of a weighted combination of the object distance map with the object skeleton. Ultimately, these methods are limited by the fact that regression is poorly suited for controlling the number of local maxima per object (S. Wolf et al., 2018). Better seed finding mechanisms were necessary to improve segmentation results.

HoverNet (Graham et al., 2019) separated the horizontal and vertical components of the distance map to the object centroid, instead of the object boundaries. The authors then used the sobel operator to find the gradients for the individual distance components, and hypothesised that these would be greatest at object boundaries and lowest at object centres, allowing for a different mechanism to select seeds. While this approach leads to many of the same issues mentioned earlier regarding to seed selection, it paved the way for the use of gradients as an effective training target.

The highly influential Cellpose method (Stringer, Wang, Michaelos, & Pachitariu, 2021) pushed this idea further and developed a training paradigm that circumvented the need for seeds altogether. In Cellpose, the training target is a flow map, which is the gradient of the horizontal and vertical distance maps. Simple numerical integration methods such as Euler integration can iteratively track these gradients, which should have a single convergence point for each instance. Remarkably, this method is close

to being *end-to-end* should the Euler integration step be incorporated in the loss calculation. Instead, likely for efficiency reasons, Cellpose pre-computes *ideal* flows using a heat diffusion algorithm, and uses these as a training target using L2 loss. In a follow up method Omnipose (Cutler et al., 2022), it was noted that, in practice, Cellpose Euler integration tended to converge to a number of disconnected clusters in elongated or non-convex instances. The authors therefore modified the distance fields to point to the object boundaries instead of the object centre, allowing the Euler integration to converge to a single continuous object skeleton.

While substantial progress has been made over the last decade to improve the alignment between the training target and the segmentation performance metric, none of the *proposal-free* methods described above are *end-to-end*, as they all rely on surrogate training objectives. The drawback is that they rely on a number of hyperparameters that govern complex postprocessing functions, such as seed selection or the number and size of the steps in Euler integration. These parameters cannot be optimised using gradient descent and often lead to segmentation mistakes when models outputs are not confident, boundaries between object are ambiguous, or when objects have atypical shapes and sizes. The other drawback of these methods is that the connected component, watershed or flow tracking algorithms are all iterative algorithms that cannot be efficiently accelerated on modern GPUs that rely on parallelisation for speed up. As a result, postprocessing steps can end up taking an order of magnitude longer than applying a model (Greenwald et al., 2022), (Stringer et al., 2021). Alternative training paradigms are needed.

2.4.2 Proposal-based methods

Another branch of instance segmentation methods, derived from the broader object-detection field, was being developed in parallel to the proposal-free methods. These methods are typically two-staged: a first stage proposes bounding boxes or object centroids, and a second stage predicts a dense binary object for each proposal, hence the term *proposal-based*.

The first object detectors were based on sliding window mechanisms paired with classifiers could recognise handwritten digits (LeCun et al., 1989). Later, object detectors such as Single Shot Detectors (SSD) (W. Liu et al., 2016) used convolutional networks to regress the bounding box coordinates of objects. These convolutional networks output a number of feature maps Z_s at S spatial resolutions. At each tensor indices

i, j in each Z_s, the model predicts four bounding box coordinates as well as logits corresponding to object classes, which include a "no-object" class. These models use a regression loss such as L2 loss for coordinate prediction and a classification loss such as cross-entropy for class prediction. Early versions of these methods suffered from very low inference speeds due to the huge number of potential proposals that had to be considered. A popular variation was introduced in R-CNN (Girshick, Donahue, Darrell, & Malik, 2014), which split the proposal stage into two steps, a first stage which filtered out most of the negative examples, and a second stage that further screened the remaining proposals. Future iterations mostly focused on improving the computational efficiency by parallelising both these two steps and included Fast-RCNN (Girshick, 2015) and Faster-RCNN (Ren, He, Girshick, & Sun, 2016). These early methods focused solely on bounding box regression and while they enabled the detection of objects, they did not immediately allow for instance segmentation.

Following the success of the R-CNN approaches to bounding box regression in natural images, the Mask R-CNN method was developed to enable the prediction of dense maps for each candidate proposal which meant that the method could be used for instance segmentation (He, Gkioxari, Dollár, & Girshick, 2018). During training, Mask R-CNN optimises three losses simultaneously: bounding-box regression, bounding-box classification, and binary mask segmentation within each box. However, these models can struggle to differentiate touching objects that have similar bounding boxes. The multiple stages in the prediction also reduced inference speeds and involved complex post-processing operations to merge redundant proposals that could not be directly optimised by gradient descent.

A popular improvement over bounding box detection methods in the field of cell segmentation is the StarDist algorithm (Schmidt et al., 2018). The authors note how Mask-RCNN struggles to segregate objects that have similar bounding boxes, such as closely touching cells. Instead, for each foreground pixel, StarDist predicts the distance to the closest object boundary along a large number (e.g. 32) of radial directions. This means that the method directly predicts a polygonal outline of each object, which is often a good approximation of the object mask. This has the huge advantage of not requiring additional steps to recover objects masks, which simplifies training and accelerates inference. However, because the authors use Non Maximal

Suppression (NMS) to keep only one polygon per object, this method is fundamentally limited to approximating objects as star-convex polygons. Interestingly, a different polygon merging strategy (such as taking the union of overlapping polygons) could alleviate this limitation, at the risk of merging neighbouring objects.

As the field of instance segmentation in natural images progressed, the focus gradually shifted to resolving ambiguity in object detection. Many objects in natural images contain sub-components (e.g. a shirt vs a person wearing it) and recent methods focus on resolving this ambiguity by using prompt-segmentation. A popular example of prompt-segmentation is the Segment Anything Model (SAM) (Kirillov et al., 2023) which was released as a foundational model for universal instance segmentation and uses a vision transformer as an image encoder. While the model relies on user prompts such as bounding box or centroids, the authors suggest a crude solution for automatic segmentation: using a uniform grid of points across the image. In the microscopy domain, this method has shown to be a useful tool for semi-automatic annotation of images, but very slow inference speeds and poor automatic prompt generation limits its widespread use, especially in large microscopy images. In general, the SAM sampling approach is poorly suited to microscopy images given the wide variation of object densities.

2.4.3 Embedding-based methods

A third family of instance segmentation methods, called the embedding-based methods, seek to combine the strengths of both proposal-based and proposal-free methods. These methods learn an embedding space in which pixels belonging to the same object are mapped close together, while pixels from different objects are pushed apart. Representing objects in this way offers several advantages: the similarity between any two pixels can be directly optimized using gradient descent (Fathi et al., 2017). Some embedding based methods, such as Neven, Brabandere, Proesmans, and Van Gool (2019), use binary classification to link pixels that belong to the same instance, which facilitates postprocessing. After training, pixels with similar embeddings can be grouped together to recover the objects using a clustering algorithm such as mean-shift (De Brabandere, Neven, & Van Gool, 2017).

Embedding-based methods hence behave somewhat like proposal-based methods, because, conditioned on any foreground pixel, they can generate a binary mask of the object containing that pixel. At the same time, they retain the one-step prediction advantage of proposal-free methods, making them well suited for the segmentation of densely packed objects.

Despite these advantages, few embedding-based methods have made their way into biological imaging workflows. This is due in part to a lack of open-source implementations, but also due to limitations of methods that are available. These include restrictive assumptions on the distribution of embeddings that belong to the same object, which limits the complexity of object shapes that can be segmented. The computational overhead associated with pixel clustering can be slower than other proposal-free methods. Consequently, their routine adoption in biology will likely depend on the development of faster, more flexible implementations.

2.4.4 Comparison and state-of-the-art for cell and nucleus detection

The development of automated cell detection algorithms in microscopy images is a deceptively difficult task and research in the field has been ongoing for more than half a century (Meijering, 2012). As with many topics in computer vision, new cell segmentation algorithms are being published regularly. This makes it increasingly challenging to assess progress or determine whether any method can truly be considered the *best* for the task.

Public competitions offer a practical way to assess progress. In the 2018 Data Science Bowl (DSB) (Caicedo et al., 2019), deep-learning methods decisively outperformed traditional segmentation approaches, with U-Net–based proposal-free methods surpassing bounding-box approaches such as Mask R-CNN. In the more recent NeurIPS CellSeg challenge (Ma et al., 2024), the leaderboard was dominated by transformer- and U-Net–based architectures. The second-place team combined StarDist and HoverNet objectives. The winning entry employed a vision transformer with the Cellpose training objective and postprocessing, though a later study suggested that the original U-Net–based Cellpose might have won under the same conditions (Stringer & Pachitariu, 2024). Overall, competition results consistently favoured proposal-free approaches, in contrast to the dominance of proposal-based methods in natural image segmentation.

While competition results are useful indicators of algorithmic performance, they are not generally effective method comparators (Maier-Hein et al., 2018). In practice, few of trained models that won the DSB and NeurIPS CellSeg competitions have been adopted in daily biology workflows. Arguably, the most impactful and useful segmentation methods are those with user-friendly interfaces that are actually used by biologists.

2.5 End to end methods for bioimage analysis

Bioimage analysis workflows often take the form of pipelines which can involve a number of algorithms, methods and software. For example, a typical workflow can include image registration, denoising, cell segmentation, cell clustering, spatial analysis and micro-environment identification followed by statistical analyses to test a biological hypothesis (Jiménez-Sánchez et al., 2022). In effect, due to the complexity of imaging data and the biological mechanisms they capture, bioimage analysis tools rarely function in isolation and often rely on an expanding ecosystem of methods. To be usable by biologists, these methods have to be interoperable, operate on a range of imaging formats and be able to export results for downstream analyses. A solution to addressing this complexity has been the development of software.

2.5.1 A note on the usability of bioimaging methods

The simultaneous rise in complexity of imaging data and the tools to analyse it presents a new major challenge for the bioimaging community, one of usability and interoperability. A number of publicly funded bioimage analysis methods end up only being used by the research groups that developed them, resulting in repetitive work and an ineffective use of public research funds. Implementing and maintaining usable and high-impact bioimage computational tools can be challenging and time consuming, and is often rewarded unequally compared to developing novel algorithms in academia. On one side, computer scientists tend to prioritise novelty and efficiency over usability, and on the other, biologists typically favour research that test biological hypotheses. This environment favours proof-of-principle papers describing methodologies that seldom advance our understanding of biology (Carpenter, Kamentsky, & Eliceiri, 2012).

Usable computational methods and software tend to emerge from tight collaborations between users and developers, which ensures that the tools are up to date with state-of-the-art methods while maintaining real-world relevance (Carpenter et al., 2012). Of note, the authors stress the importance of user-friendliness, modulation, validation and interoperability. Of particular importance is the open-source label, which ensures broad accessibility and adaptability through adherence to the Open Source Definition (OSD) (https://opensource.org).

The OSD goes beyond availability of source code. To qualify as open-source, software must permit both commercial and non-commercial use, among other criteria. In this regard, the label is often misused within the bioimage analysis community, especially among groups that share foundational models trained on a wide range of publicly available datasets with restrictive non-commercial or research-only licenses.

The research presented in this thesis is academic, and hence focuses on novelty and originality. Nonetheless, in pursuit of a wider impact within the community, we emphasise both usability and openness to the greatest extent feasible.

ImageJ and Fiji

The personal computer revolution and its adoption in research settings was immediately accompanied with a need for image visualisation and analysis tools. One of the first software to fill this demand was the Pascal-based NIH Image, a precursor of the now popular Java-based ImageJ, written by Wayne Rasband (Schneider, Rasband, & Eliceiri, 2012). Freely distributable, interpretable and well-documented code encouraged the community development of plugins and macros which allowed for a growing functionality base, in turn enabling a growing application base (Schneider et al., 2012). While ImageJ was written by biologists for biologists, the software eventually drew interest from the computer science community which now maintain and develop ImageJ bundled with a large array of plugins under the form of Fiji (Fiji Is Just ImageJ) (Schindelin et al., 2012). To this day, ImageJ and Fiji are by far the most widely used and cited image analysis software used by the biology community, and has inspired a range of open-source software.

QuPath

A popular Java-based software inspired by the open-source and modular ImageJ template is QuPath (Bankhead et al., 2017). QuPath was originally developed for the visualisation, annotation and analysis of very large 2D microscopy images without requiring high performance hardware. QuPath's main contributions are the efficient handling of objects, detections and annotations as well as real-time training of pixel and object classifiers using conventional machine learning techniques. Although the traditional user-base was the digital pathology community, QuPath allows for scripting to extend the functionality and scope of the software and this has greatly broadened its user base. QuPath's interoperability with other imaging platforms (e.g. OMERO, ImageJ), file formats and a range of community developed plugins has enabled it to remain highly relevant amidst the increasing popularity of Python-based deep-learning methods.

The Python ecosystem

Over the last decade, Python has solidified its position as the leading programming platform for exploratory research and image analysis development. This is primarily due to its extensive collection of libraries, such as the general multidimensional-array handling library NumPy (Harris et al., 2020) and specialised image processing libraries Scikit-Image (Van der Walt et al., 2014). The remarkable rise in popularity of deep-learning libraries such as Tensorflow (Abadi et al., 2016) and PyTorch (Paszke, 2019) has made Python nearly indispensable for developing and training deep-learning models. Furthermore, with extensive GPU support, PyTorch can also greatly accelerate the handling of multidimensional arrays for image processing.

Nevertheless, the ease of deployment and installation of Java-based software has made tools written in this language preferred among many biologists and clinicians who may lack the technical expertise or time to install Python and its libraries. Yet, the simplicity of Python as a programming language has allowed it to dominate exploratory and experimental computing research. To increase the usability of modern computer vision algorithms for biologists, successful solutions have been (1) creating more computational bridges between the two programming languages (e.g. Jython, TorchScript, ONNX), and (2) lowering the entry barrier to the usage of Python-based image analysis platforms, such as the interactive multidimensional-image visualisation and analysis platform Napari (Sofroniew et al., 2025).

2.5.2 Bridging the computational gap between computer science and biological research

Computational bridges and standardisation between computing platforms are crucial for improving the accessibility, usability and reproducibility in image analysis. State-of-the-art computer vision methods, especially those that rely on deep-learning models, are developed in settings that differ in terms of hardware, operating system and programming language to the settings typically used by researchers during deployment. Substantial effort has been made by the open-source community to help the transition to deployment, including the Open Neural Network Exchange (ONNX) (Bai, Lu, Zhang, et al., 2019) and TorchScript (Paszke, 2019) for more complex models (DeVito et al., 2021). However, in the bioimage analysis setting, these efforts are often undermined by the complexity of computational steps that are performed before or after the application of the model.

A big step towards the democratisation of the latest AI developments in the life sciences is the BioImage Model Zoo (Ouyang et al., 2022), which provides a platform for sharing and reusing pre-trained models with standardised metadata for diverse biological use cases, including pre and post-processing pipelines. To date, The Bioimage Model zoo hosts models for diverse applications including image classification, super-resolution, denoising, instance segmentation, semantic segmentation and image-to-image translation (Franco-Barranco et al., 2024).

2.6 Aims and outline of this thesis

The aim of this thesis is to address current challenges in cell and tissue phenotyping by developing and validating computational methods for the accurate, efficient and accessible analysis of microscopy images. We illustrate the main methods that we develop in this thesis in Fig. 2.5.

The work presented herein is structured as follows:

Chapter 3 introduces InstanSeg, a novel instance segmentation algorithm that learns a pixel embedding space optimised for efficient object separation. We demonstrate that our approach not only achieves state-of-the-art accuracy on public histology datasets, but also addresses computational limitations in processing speed. InstanSeg's

lightweight tensor-based postprocessing enables high-throughput analysis of whole slide images and large image collections, as well as enabling integration into the user-friendly graphical software QuPath, lowering the barrier to adoption for the wider biological community.

Chapter 4 extends InstanSeg for the segmentation of cells in highly multiplexed fluorescence imaging. We first identify limitations of current deep-learning based architectures for the processing of images with varying number and ordering of imaging channels. To address this, we present ChannelNet, a novel architecture designed to produce fixed representations of images with arbitrary arrangement of imaged biomarkers. By coupling ChannelNet with InstanSeg, we achieve state-of-the-art performance for the joint segmentation of nuclei and cells, importantly we show that segmentation performance improves as the number of biomarker channels increases.

Chapter 5 addresses the downstream task of cell classification, demonstrating that cell phenotyping can be effectively decoupled from instance segmentation. For this, we leverage registered immunohistochemistry images to generate a large number of weak annotations to train large deep-learning classifier for the detection of immune cell types in brightfield images of renal biopsies. Our method achieved joint-first place in the MONKEY public competition hosted on Grand Challenge.

Chapter 6 investigates the effect of cell segmentation on downstream representation of cell features in multiplexed fluorescence images. We also introduce a new architecture for the phenotyping of cell populations based on cell-wise biomarker expression. We train this model using purely synthetic data and show high generalisation performance on real images.

We conclude the thesis by briefly discussing the adoption of these tools within the biological community and mention some of the current biological and clinical applications of InstanSeg.

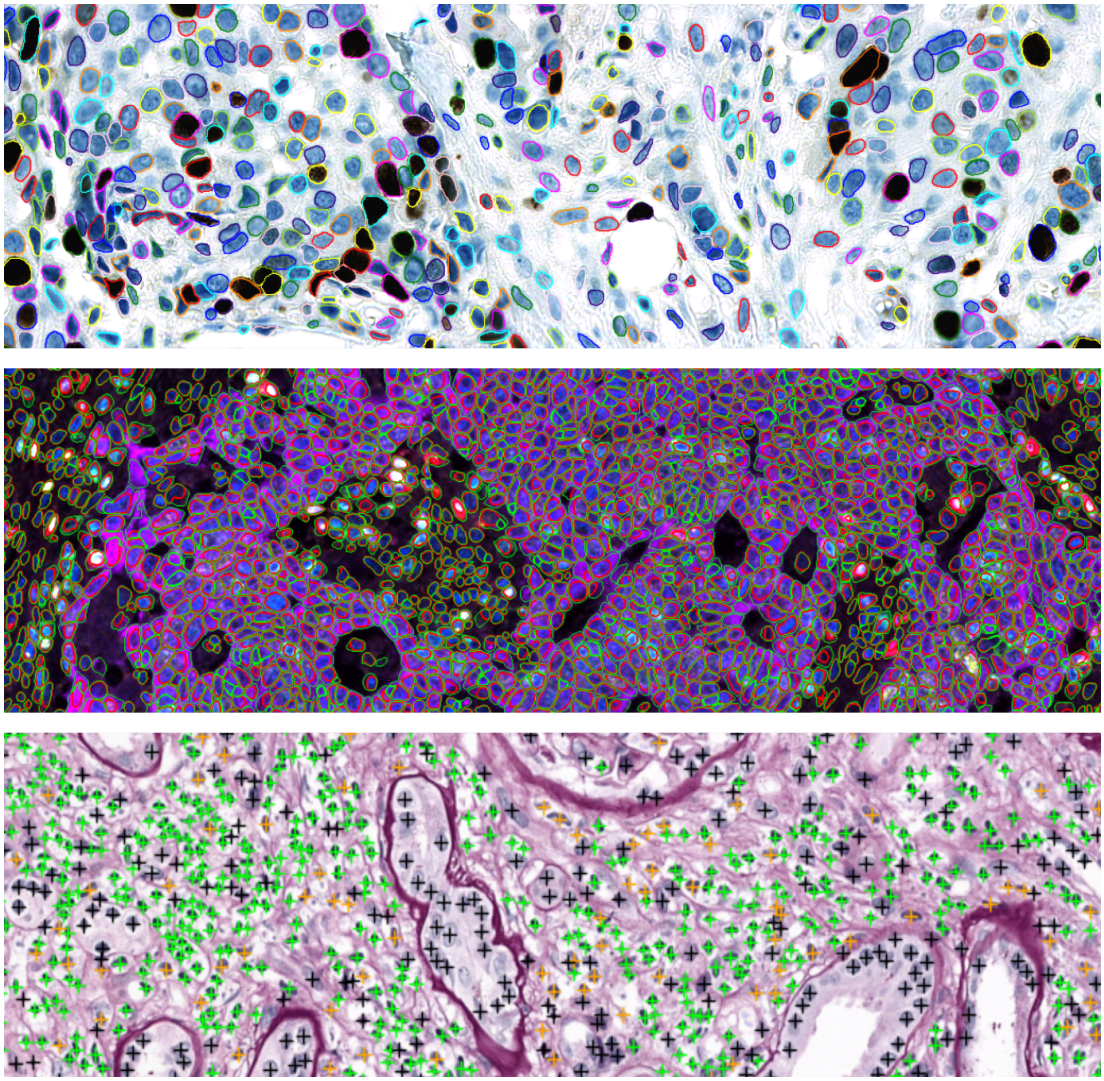

Figure 2.5: Summary figure of some of the algorithms presented in this thesis. From top to bottom, an efficient algorithm for the segmentation of nuclei in brightfield images (OS-2.ndpi, OpenSlide), a novel method for the joint segmentation of nuclei (green) and cells (red) in multiplexed fluorescence images (CPDMI, Aleynick et al. (2023)). An algorithm for the detection and classification of immune cell types in brightfield images (MONKEY, Midden et al. (2024))

Chapter 3

InstanSeg: an embedding-based instance segmentation algorithm optimized for accurate, efficient and portable cell segmentation

3.1 Abstract

Cell and nucleus segmentation are fundamental tasks for quantitative bioimage analysis. Despite progress in recent years, biologists and other domain experts still require novel algorithms to handle increasingly large and complex real-world datasets. These algorithms must not only achieve state-of-the-art accuracy, but also be optimized for efficiency, portability, and user-friendliness. Here, we introduce InstanSeg: a novel embedding-based instance segmentation pipeline designed to identify cells and nuclei in microscopy images. Using six public cell segmentation datasets, we demonstrate that InstanSeg can significantly improve accuracy when compared to the most widely used alternative methods, while reducing the processing time by at least 60%. Furthermore, InstanSeg is designed to be fully serializable as TorchScript and supports GPU acceleration on a range of hardware. We provide an open-source implementation of InstanSeg in Python, in addition to a user-friendly, interactive QuPath extension for inference written in Java. Our code and pre-trained models are available at github.com/instanseg/instanseg.

3.2 Introduction

Quantitative cell biology and digital pathology frequently involve extracting and evaluating cell features from imaging data. This includes the size, shape, location and staining of cells, as well as tissue-level properties such as cell counts and interactions. Determining features from individual cells relies on accurate detection and boundary identification. Detecting stained nuclei is typically used as a proxy for segmenting entire cells, especially in brightfield or fluorescence images where cell boundaries may not be visible. Despite its ubiquity, the task of identifying nucleus boundaries from microscopy images has proven to be deceptively difficult (Meijering, 2012), and remains a very active area of research, including grand challenges and contests (Caicedo et al., 2019). The difficulty of the task arises from the immense variation seen in biological images combined with the specific problem of distinguishing nuclei that may be densely packed and, therefore, have indistinct boundaries. The overall challenge fits into the more general computer vision field of instance segmentation.

The goal of instance segmentation is to assign pixels to individual instances. Most current methods can be categorized as proposal-based or proposal-free methods. Proposal-based methods, such as Mask-R-CNN (He et al., 2018), Fast-R-CNN (Girshick, 2015) or more recently SAM (Kirillov et al., 2023), typically rely on a Deep Neural Network (DNN) to predict bounding boxes for every instance, followed by a second network that predicts a dense binary mask for every bounding box. These two-step approaches typically come at a high computational cost and slower inference speeds, and are seldom used in end-to-end implementations by biologists. An exception to this is the popular Stardist (Schmidt et al., 2018) method, which relies on the prediction of a more detailed, usually 32-sided bounding polygon. This circumvents the need for the second-step binary mask prediction, at the cost of some fine-grained detail at the object boundary.

The alternative to these approaches are the proposal-free methods, which usually rely on a Fully Convolutional Network (FCN) predicting dense feature maps that are later post-processed to resolve individual instances. Segregating foreground from background pixels is relatively standard across methods and relies on the prediction of a foreground probability map that can be thresholded. The main differences between current methods relate to how they separate touching or overlapping instances. One popular approach relies on the prediction of an instance-boundary distance map (Greenwald et al., 2022; Naylor, Lae, Reyal, & Walter, 2019). Other methods use

two dimensional offsets (Graham et al., 2019) or flows (Stringer et al., 2021) to the instance centroid. These serve as input to a watershed or flow tracking algorithm, which is usually seeded at local maxima in the foreground probability map. In practice, it is difficult for any predictor to place exactly one seed for every instance, which is crucial to avoid errors in the segmented output. Furthermore, the computational cost of a watershed transform commonly causes a performance bottleneck. Especially in the field of computational pathology, where a single whole-slide image (WSI) is typically 10-40 GB in size and may depict over a million cell instances, computational efficiency is a crucial practical consideration.

More recent work by Neven et al. (2019) and its adaptation to 2D and 3D microscopy images as *EmbedSeg* (Lalit, Tomancak, & Jug, 2022), introduces a novel and powerful approach to proposal-free, boundary-detailed instance segmentation. The method builds on an Erfnet (Romera, Álvarez, Bergasa, & Arroyo, 2018) backbone to generate dense pixel embeddings that are further used to cluster pixels of the same instance while segregating pixels from neighbouring instances. Despite the potential of embedding-based segmentation methods, EmbedSeg has received relatively little attention in the field. This is likely due to (1) a dependence on the Erfnet backbone, (2) non-commercial restrictions on code reuse, (3) postprocessing steps that cannot easily be ported outside of the python environment, (4) a seed sampling strategy that differs between train and test time, (5) a restrictive assumption on the distribution of embeddings around sampled seeds. Together, these hinder the accuracy, efficiency and portability of the method, and have prevented its integration both in commercial applications and in user-friendly open-source software packages widely used by biologists, such as Fiji (Schindelin et al., 2012) or QuPath (Bankhead et al., 2017).

To address these issues, we present *InstanSeg*, a novel embedding-based instance segmentation method optimized for accuracy, efficiency, and portability. Building on a modified U-Net backbone (Ronneberger et al., 2015), InstanSeg uses a lightweight neural network to cluster pixel embeddings around predicted seed locations. Our novel approach sets a new state-of-the-art in terms of accuracy on six public nucleus segmentation datasets, while reducing the processing time by a factor of approximately 2.5 – 45x compared to the most widely used current methods. Furthermore, InstanSeg is a highly vectorized pipeline that can be efficiently used for inference

on a laptop GPU and serialized in a self-contained TorchScript implementation. This strategy means that InstanSeg can be used not only in Python, but also integrated into software tools written in other languages, without a requirement to replicate any complex pre- and post-processing steps.

3.3 InstanSeg, a novel embedding-based segmentation algorithm

3.3.1 Problem setting

Our goal is to learn an instance segmentation model that takes an input image $\boldsymbol{X}$ of shape $C \times H \times W$ and predicts a labelled segmentation map $\hat{\boldsymbol{L}}$ of shape $H \times W$ which not only distinguishes instances from background but also individual neighbouring instances from each other. C denotes the number of image channels. Given a dataset $\mathcal{D}$ with $|D|$ doublets, each including an image $\boldsymbol{X} \in \mathbb{R}^{C \times H \times W}$ and an instance segmentation mask $\boldsymbol{L} \in \mathbb{R}^{H \times W}$, we would like to train the parameters of our model to estimate the ground truth $\boldsymbol{L}$. Each element of $\boldsymbol{L}$ is a discrete value between 1 and K indicating that the corresponding pixel belongs to one of K instances in the image.

3.3.2 Intuition

We treat the task as a pixel assignment problem, where we want to associate each pixel $\mathbf{X}_{ij}$ at the coordinates (i, j) to its corresponding object. As a proxy, we seek a set of *seed* pixels (e.g. centroids) to represent each object, and then relate each pixel to these seeds. A simple approach is to use a similarity metric to relate pixel embeddings to seed embeddings. However, this approach runs into two problems: (1) fully convolutional networks are translationally invariant, and may struggle to segregate pixels that are situated far apart on the image plane based on embeddings alone. (2) A similarity metric does not obviously relate to a probability of belonging to an object, especially when considering objects that are overlapping or of vastly different sizes.

The first problem can be solved by introducing a pixel coordinate system either inside the model (Kulikov & Lempitsky, 2020) or to the model outputs (Neven et al., 2019); we favor the latter for ease of implementation. For the second problem, we need to find a function that maps a similarity metric to a probability of belonging to an object. Ideally, the function should be conditioned on higher-order object properties, such as size,

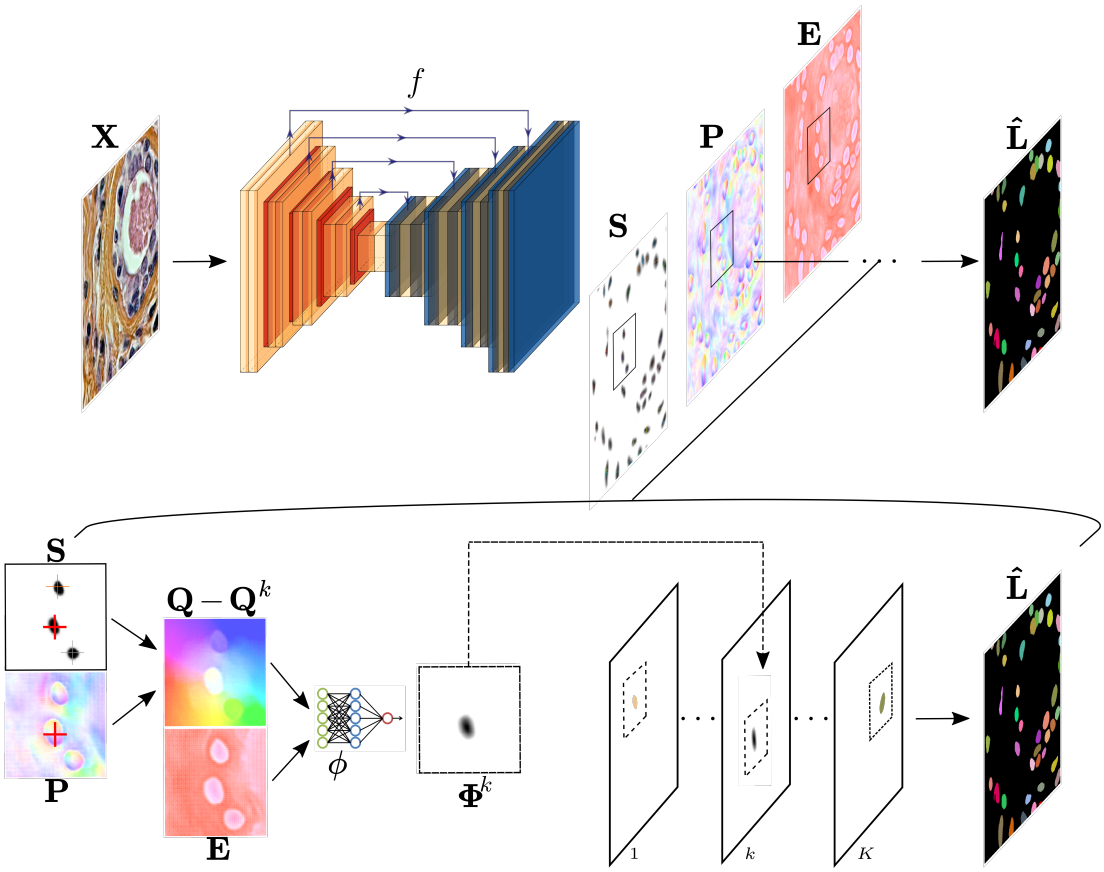

Figure 3.1: Diagram of the proposed embedding-based instance segmentation method. A feature encoder f and three feature heads s, p and e transform an input image $\boldsymbol{X}$ into a seed map $\boldsymbol{S}$, a positional embedding $\boldsymbol{P}$ and a conditional embedding $\boldsymbol{E}$. We sample local maxima in the seed map to find the coordinates of seed pixels $\boldsymbol{u}^k$ (e.g red cross) and compute the relative offsets between the positional embeddings and each seed embedding $\boldsymbol{Q} - \boldsymbol{Q}^k$. These offsets, along with conditional embeddings $\boldsymbol{e}$ serve as input to the instance segmentation head Φ, which outputs a probability map $\boldsymbol{\Phi}^k$ of each pixel belonging to instance k. The final labelled segmentation map $\hat{\boldsymbol{L}}$ is obtained by merging the K probability maps. Input image is from the TNBC 2018 dataset (Naylor et al., 2019).

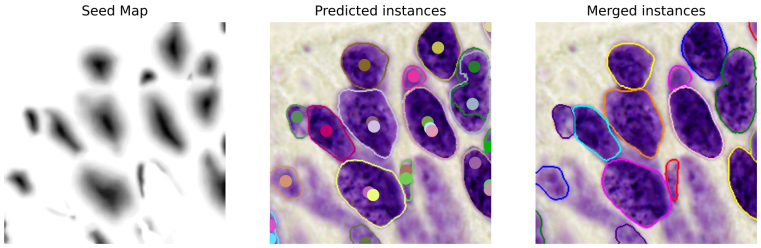

Figure 3.2: InstanSeg uses a local maxima algorithm on a predicted seed map (left image) to locate suitable seed(s) for each object. In large instances, there may be multiple seeds detected within the object. InstanSeg was trained to predict the full instance irrespective of the sampled seed location; as a result, predicted objects corresponding to the same instance greatly overlap (centre image). We merge objects with large overlaps by taking their union (right image). Input image from the DSB 2018 dataset (Caicedo et al., 2019).

orientation or shape. Previous work by Neven et al. (2019) uses a Gaussian function to map a similarity metric to a probability of belonging to an instance. While intuitive, this formulation enforces pixel embeddings to lie in a circular or elliptical distribution around the instance seed, a task that can be difficult when segmenting touching non-convex objects that do not have obvious centres. Instead, we do not seek a specific functional form, and opt to instantiate it as a neural network instead, thereby relaxing the underlying assumptions on the distribution of embeddings around instance seeds.

Hence, we solve the task in three steps, we (1) seek suitable seed pixels to represent each object, (2) compute a similarity metric to relate each pixel to each seed pixel, (3) use a neural network to map a similarity metric, conditioned on learnt higher object properties, to a probability of belonging to an object.

3.3.3 Our method

As illustrated in Fig. 3.1, we build our model with a feature encoder f and three auxiliary prediction heads: s, p and e predicting the location of seed pixels $\boldsymbol{S}$, positional embeddings $\boldsymbol{P}$ and conditional embeddings $\boldsymbol{E}$ respectively. A fourth prediction head ϕ is finally used to predict instance labels. The feature encoder f takes in a C-dimensional $H \times W$ input image and outputs a feature map $\boldsymbol{F}$ consisting of D_f feature maps, each with $H \times W$ spatial resolution. The auxiliary encoders p and e take in the feature map $\boldsymbol{F}$ and output D_p and D_e feature maps with $H \times W$ spatial dimensions respectively.

Seed selection

The aim of the seed head is to predict the location of suitable seed pixels that best represent an instance. Previous embedding-based segmentation methods have used the centroid (Neven et al., 2019) or medoid (Lalit et al., 2022) of the object's positional embeddings as their seed location. While sensible, these are not obvious to sample at test time when objects are not yet resolved – requiring a sampling strategy that differs between the training and testing phases. They also require an additional term in the loss computation.

We seek a simpler and more consistent seed sampling strategy. We hypothesise that pixels that lie close to the instance centre are suitable seed pixels and we train the seed head to predict the distance to the instance boundary for each foreground pixel. Hence, we minimize the following loss function:

$$\mathcal{L}_s = \frac{1}{|D|} \sum_{(\boldsymbol{X},\boldsymbol{L})\in\mathcal{D}} \sum_{i=1}^{H}\sum_{j=1}^{W} \ell_1(\boldsymbol{S}_{ij}, d(\boldsymbol{L}_{ij})) \tag{3.1}$$

where the subscript i, j denotes the feature at the spatial location (i, j), $d(\boldsymbol{L}_{ij})$ denotes the relative distance from the pixel location (i, j) to the nearest instance boundary, and ℓ_1 is the $L1$ loss function.

Importantly, we use the same seed sampling strategy during both training and testing: we find the location $\boldsymbol{u}^k$ of the seed pixel belonging to instance k by sampling local maxima in the seed map $\boldsymbol{S}$. Note that there may be more than one seed pixel per instance, which can easily be resolved in postprocessing. This approach allows for larger instances to be segmented by merging smaller overlapping fragments (see top-right nucleus in Fig. 3.2 for an example).

Similarity metric computation

We use the computed coordinate for each seed to obtain a seed embedding by using the auxiliary positional encoder p, *i.e.* $\boldsymbol{P}^k = \boldsymbol{P}_{\boldsymbol{u}^k}$ where $\boldsymbol{P}_{\boldsymbol{u}^k}$ denotes D_p dimensional encoding at the location $\boldsymbol{u}^k$. We then compare each embedding pixel $\boldsymbol{P}_{ij}$ to each sampled seed embedding $\boldsymbol{P}^k$ to relate pixels to their corresponding objects. To this end, we calculate the offset between $\boldsymbol{P}_{ij}$ and $\boldsymbol{P}^k$. Specifically, we compute $\boldsymbol{Q}_{ij} - \boldsymbol{Q}^k = (\boldsymbol{P}_{ij} + \boldsymbol{O}_{ij}) - (\boldsymbol{P}^k + \boldsymbol{O}_{\boldsymbol{u}^k})$ where $\boldsymbol{O}$ denotes linear coordinates in the H and W dimensions. Unlike previous embedding-based methods, we do not limit the

dimensionality of the coordinate system $\boldsymbol{O}$ to match the input image dimensionality – specifically, when $D_e > 2$, we match the dimension of $\boldsymbol{O}$ by adding empty channels, allowing for embeddings $\boldsymbol{Q}_{ij}$ to lie outside of the image plane. We hypothesize that the increased embedding space enables better separation of crowded instances.

Instance probability mapping using a neural network

For each seed embedding $\boldsymbol{P}^k$, we use a separate head ϕ to predict a binary probability map $\boldsymbol{\Phi}^k$ of each pixel belonging to the instance in which $\boldsymbol{u}^k$ resides. A simple segmentation head could hence take pixel offsets $\boldsymbol{Q}_{ij} - \boldsymbol{Q}^k$ as sole input. However, learnt pixel offsets may not be sufficient to separate overlapping or crowded instances. For example, consider a small instance bordering a much larger one, pixel offsets near their boundary will tend to be closer to the smaller instance's seed and lead to a biased pixel assignment. To address this, we use pixel embeddings $\boldsymbol{E}_{ij}$ from a separate encoder e as an additional input to the segmentation head, and hypothesize that these can encode higher order object properties such as orientation, shape or size.

The instance segmentation head $\phi : \mathbb{R}^{D_e+D_p} \to \mathbb{R}^1$ takes in $\boldsymbol{Q}_{ij} - \boldsymbol{Q}^k$ along with $\boldsymbol{E}_{ij}$ and outputs a scalar $\boldsymbol{\Phi}^k_{ij}$. The head ϕ is instantiated as a multi-layer perceptron (MLP) with a single hidden layer and is trained by minimizing the loss

$$\mathcal{L}_i = \frac{1}{|D|} \sum_{(\boldsymbol{X},\boldsymbol{L})\in\mathcal{D}} \sum_{k=1}^{K} \sum_{i=1}^{H} \sum_{j=1}^{W} \ell_{lh}(\boldsymbol{\Phi}^k_{ij}, \boldsymbol{L}^k_{ij}), \tag{3.2}$$

where ℓ_{lh} is the Lovasz-hinge loss (Berman et al., 2018), and the scalar $\boldsymbol{L}^k_{ij}$ is given by

$$\boldsymbol{L}^k_{ij} = \begin{cases} 1 & \text{if } \boldsymbol{L}_{ij} = k \\ 0 & \text{if } \boldsymbol{L}_{ij} \neq k. \end{cases}$$

.

While the Lovasz-hinge loss has a higher complexity $O(n \log n)$ than per-pixel losses such as cross-entropy or hinge loss $O(n)$, the Lovasz extension directly optimises a convex surrogate of the Jaccard index, often leading to better results when assessed using segmentation metrics that indirectly rely on Jaccard, such as the F1 score.

In effect, our instance segmentation conditions pixel embeddings on seed embeddings using a simple addition operation, this bears resemblance with the additive attention mechanism introduced in Bahdanau, Cho, and Bengio (2014). Under this formulation, each seed embedding acts as a key and each pixel embedding as a query.

Since the instance and seed losses had similar scales, we used an unweighted sum as the final loss function:

$$\mathcal{L} = \mathcal{L}_s + \mathcal{L}_i.$$

In the interest of computational efficiency, we only consider pixels in the vicinity of the sampled seed locations, as illustrated in the pipeline diagram. Specifically, we compute a square window around each seed location, with the size of the window constraining the maximal size of a recoverable object. As a result, the memory requirement of postprocessing the model outputs $\boldsymbol{F}$ to a labelled segmentation map $\hat{\boldsymbol{L}}$ scales linearly with the number of instances, independently of image size.

Inference

We use an identical seed sampling strategy as in our training phase. There may be multiple local maxima sampled within a single object; this is difficult to avoid, as in general it is hard for any predictor to place exactly one seed inside each object. However, the network ϕ was trained to output the same probability map $\boldsymbol{\Phi}^k$ for any seed embedding $\boldsymbol{P}^k$ that was sampled within the instance k. As a result, we can identify redundant probability maps $\boldsymbol{\Phi}^k$ using an intersection-over-union metric. We merge redundant $\boldsymbol{\Phi}^k$ by taking their union, see Fig. 3.2 for an illustration. Finally, we obtain a labelled segmentation map using

$$\hat{\mathbf{L}}_{ij} = \begin{cases} \arg\max_{(k)} \boldsymbol{\Phi}^k_{ij} & \text{if } \max_{(k)} \boldsymbol{\Phi}^k_{ij} \geq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{3.3}$$

Table 3.1: Total number of images and instances for each dataset.

Dataset	Images (Train/Val/Test)	Instances (Train / Val / Test)
CoNSeP (Graham et al., 2019)	21 / 6 / 14	12640 / 2915 / 8777
TNBC 2018 (Jack, Tsai, Marick, & Yamada, 2021)	54 / 7 / 7	4434 / 398 / 349
MoNuSeg (Kumar et al., 2017)	29 / 8 / 14	18195 / 5978 / 6699
LyNSeC (N. Hussein et al., 2023)	559 / 70 / 70	121436 / 13910 / 16433
NuInsSeg (Mahbod et al., 2024)	532 / 66 / 67	24570 / 3219 / 2909
IHC TMA (R. Wang et al., 2024)	212 / 27 / 27	7284 / 971 / 906

3.3.4 Backbone network architecture

We reviewed the accuracy and performance of a number of publicly available convolutional and transformer based model architectures (results not shown). We found the best performing architecture was the UNet (Ronneberger et al., 2015) based implementation from Cellpose (Stringer et al., 2021), which uses maxpooling in the encoder blocks and nearest neighbour interpolation in the decoder blocks, with residual connections in each block. The decoder layers use summation instead of concatenation to merge skip connections. We therefore adopt a modified version of Cellpose's backbone network. Specifically, we remove the style vectors to improve translational invariance and we change the ordering of operations inside each block to convolution - normalisation - activation, as default in other libraries. Our resulting backbone model has approximately four million parameters, due to the removal of the style vectors and associated weights, our model has around half the number of trainable parameters compared to the original Cellpose model, with similar accuracy.

3.3.5 Training details

InstanSeg is implemented using the Pytorch library. We train over 500 epochs, each consisting of 1000 batches of size 3. Each image is a random 256×256 pixel crop. We use a fixed learning rate of 0.001 and use the Adam optimizer (Kingma & Ba, 2017), we store model weights when the highest scoring F_1^μ on the validation set is reached.

We use an additional pretraining step of 10 epochs where we substitute L1 regression of the distance map with binary cross entropy in Eq. (3.1) and replace the Lovasz-hinge loss with Dice loss in Eq. (3.2). We find that this greatly accelerates and stabilizes convergence in the initial stages of training. During training, we cap the number of instances in Eq. (3.2) to $K = 50$ for lower GPU memory requirements and faster convergence.

To fairly compare InstanSeg with other methods, we do not resize the input images for the benchmark results. However, in the models we make publicly available, we use the known image resolution in terms of µm/px to standardize the scale the training images. This solves the problem of how to resize images acquired at different magnifications to be compatible with the trained model.

3.3.6 Augmentations

For the benchmarks, we use minimal augmentations during training so as to fairly compare results with other methods. These include horizontal and vertical flips, axis-aligned rotations and random crops of size 256×256 pixels. For additional models we make publicly available, we use further augmentations including stain normalization, hue, brightness and contrast shifts.

3.3.7 Tiled predictions

The GPU memory requirements of InstanSeg are low enough for each of the validation and test set images without having to tile or resize the input images, therefore we only pad the input images so that the height and width are divisible by 32 as required by our U-Net backbone. For inference on larger images, such as whole slide images, we run InstanSeg on individual tiles to obtain labelled images with a fixed overlap of 80 pixels and merge the predicted labels by matching duplicated objects using an IoU metric. This differs from other tiling approaches that merge the intermediate model outputs and postprocess the resulting stitched image. Our approach enables InstanSeg to run on WSIs that are bigger than system RAM and can benefit from GPU acceleration for postprocessing model outputs.

3.3.8 Test time augmentations

Test time augmentations (TTA) involve running a model on a set of augmented versions of a test image and aggregating the model predictions to increase model accuracy at the expense of compute time and/or memory. TTA is an expensive operation that is seldom used in real-world applications, but is commonly reported in segmentation benchmarks. To fairly compare with other methods, we report both results with and without TTA in our benchmarks. Our TTA implementation uses ttach (Iakubovskii,

Table 3.2: Quantitative segmentation results on 6 publicly available datasets. Best results are shown in bold, second best results in italics.

	TNBC 2018		NuInsSeg		MoNuSeg		IHC TMA		CoNSeP		LyNSec	
	F_1^{μ}	$F_1^{0.5}$	F_1^{μ}	$F_1^{0.5}$	F_1^{μ}	$F_1^{0.5}$	F_1^{μ}	$F_1^{0.5}$	F_1^{μ}	$F_1^{0.5}$	F_1^{μ}	$F_1^{0.5}$
StarDist	0.645	*0.896*	0.494	0.799	0.543	0.846	0.470	0.798	0.418	0.690	0.701	0.920
HoVer-Net	0.546	0.768	0.374	0.635	0.438	0.707	0.304	0.559	0.312	0.538	0.659	0.886
CellPose	0.627	0.835	0.497	0.788	0.553	0.850	0.545	0.811	0.389	0.626	0.701	0.911
EmbedSeg (TTA)	0.641	0.870	0.492	0.761	0.560	0.853	-	-	0.414	0.618	-	-
InstanSeg	**0.698**	**0.897**	*0.514*	*0.803*	**0.573**	*0.858*	*0.560*	*0.820*	*0.478*	*0.701*	**0.725**	*0.922*
InstanSeg (TTA)	*0.690*	0.880	**0.521**	**0.808**	*0.568*	**0.859**	**0.562**	**0.835**	**0.498**	**0.723**	*0.722*	**0.924**

2024), and involves 16 combinations of axis aligned rotations and flips. We cannot directly pool the model outputs, as these are orientation dependent. Instead, we merge the outputs of our segmentation head $\mathbf{\Phi}^k$ using element wise median pooling. Unless explicitly specified, we do not use TTA when reporting our results.

3.4 Baselines, experiments and results

3.4.1 Datasets

We benchmark InstanSeg on six independent publicly available datasets. We focus on datasets with clearly-defined licensing terms to facilitate reuse *TNBC 2018* (Jack et al., 2021), *NuInsSeg* (Mahbod et al., 2024), *IHC TMA* (R. Wang et al., 2024), *CoNSeP* (Graham et al., 2019), *MoNuSeg* (Kumar et al., 2017) and *LyNSeC* (N. Hussein et al., 2023). We report summary dataset statistics in Table 3.1.

We do not include the DSB 2018 dataset (Caicedo et al., 2019) for three main considerations: (1) the architecture and rationale behind building InstanSeg was extensively developed using this dataset and reporting results unfairly benefits our method over the other baselines, (2) a comparatively high number of labelling mistakes and (3) the lack of pixel resolution information, limiting the use of this dataset in real-world applications.

3.4.2 Evaluation metrics

The F_1 score was used as a metric for detection accuracy. Predicted objects having an IoU with a ground-truth object greater than the IoU threshold τ are considered true positives T_p, while if the IoU is smaller than τ, it is considered false positives F_P. The number of false negatives F_N is calculated as the difference between the number

of ground truth objects and the number of true positives. The score is calculated as $F_1 = \frac{2T_P}{2T_P+F_P+F_N}$. We report both $F_1^{0.5}$, determined at $\tau = 0.5$ and F_1^{μ}, calculated as the mean F_1 score over the interval $[0.5, 0.9]$ with a step of 0.1. Our metrics are calculated using the Stardist implementation [1].

3.4.3 Baselines

For evaluating *InstanSeg*, we selected the three most widely-used nucleus and cell detection methods for bioimage analysis based on deep learning, along with a fourth method that is the most similar to our proposed approach. These methods are widely reported to represent the current state-of-the-art, and cover a range of different segmentation approaches.

We do not report previously published results because exact datasets, training splits and evaluation metrics differ among the literature. We retrain all methods from scratch on identical train, validation and test splits on all six datasets. For all baselines, we use the official code and default hyper-parameters when possible. Computational restraints prevent us from optimizing training hyperparameters (e.g. batch size, learning rate, regularization) on the individual datasets, as some of the methods required multiple days of training. As a result we use default training parameters for all methods and datasets. We acknowledge that this might not fully exploit the potential of each method and lead to some of the methods under-performing. The default hyperparameters provided by each study were selected by the respective authors based on similar datasets and similar performance metrics, we therefore expect that these translate to the datasets used in this study. To improve fairness, we optimise InstanSeg's hyperparameters on a separate dataset (DSB 2018) and refrain from including this dataset in our results.

Cellpose (Stringer et al., 2021) is a popular whole cell and nucleus segmentation pipeline. The method is based on a modified UNet backbone trained to predict a simulated heat diffusion pattern initiated at instance centroids. Despite being computationally expensive, the method is known to achieve high accuracies for both nucleus and whole cell segmentation, and was originally intended as a generalist algorithm that performs well over a wide range of image modalities and image resolutions. We run Cellpose with default parameters, on all three (RGB) input channels.

1. https://github.com/stardist/stardist

Stardist (Schmidt et al., 2018) uses a modified UNet backbone to predict a vector for every pixel in an image. During training, the vector is regressed to the distance to the object boundary along a set of predefined angles termed radial directions. Following a non-maximal suppression step, the method predicts a star-convex polygon approximating every instance. We use N = 32 radial dimensions.

HoVer-Net (Graham et al., 2019), is an encoder-decoder network predicting horizontal and vertical distances to the instance centre of mass which are then post-processed using a marker controlled watershed on the predicted gradients. The method was originally intended to both segment and classify nuclei, but we only consider the segmentation branch of the method. For timing Hover-Net, we use the Monai implementation and set the postprocessing mode to “Fast". We use the default sliding window inference function using a batch size of 8.

EmbedSeg (Lalit et al., 2022) is a recent adaption of (Neven et al., 2019) to microscopy images and is conceptually the method most similar to the proposed InstanSeg. We set n_sigma to 5 and keep all parameters as default. Note that we were unable to produce meaningful segmentation results on the IHC TMA and LyNSeC datasets without modification of the source code. It was unclear why the method failed on these datasets, but manually changing the values of n_x and n_y parameters at test time improved results. We omit EmbedSeg results on these two datasets. We use test-time augmentations (TTA) when reporting the accuracy results, but disable TTA when reporting time efficiency of the method. When reporting time efficiency, we set the postprocessing mode to “Fast".

Results and discussion

We benchmark InstanSeg on six independent public nucleus segmentation datasets and report our results in Table 3.2 , we further display per-image F1 scores in Fig. 3.3. We find that InstanSeg’s performance is consistently superior to the previous state-of-the-art nucleus segmentation pipelines. Our base InstanSeg method surpasses all previous methods on 11/12 metrics. We report paired t-tests in Fig. 3.3 and show that our method provides statistically significant improvements on all but the IHC TMA dataset.

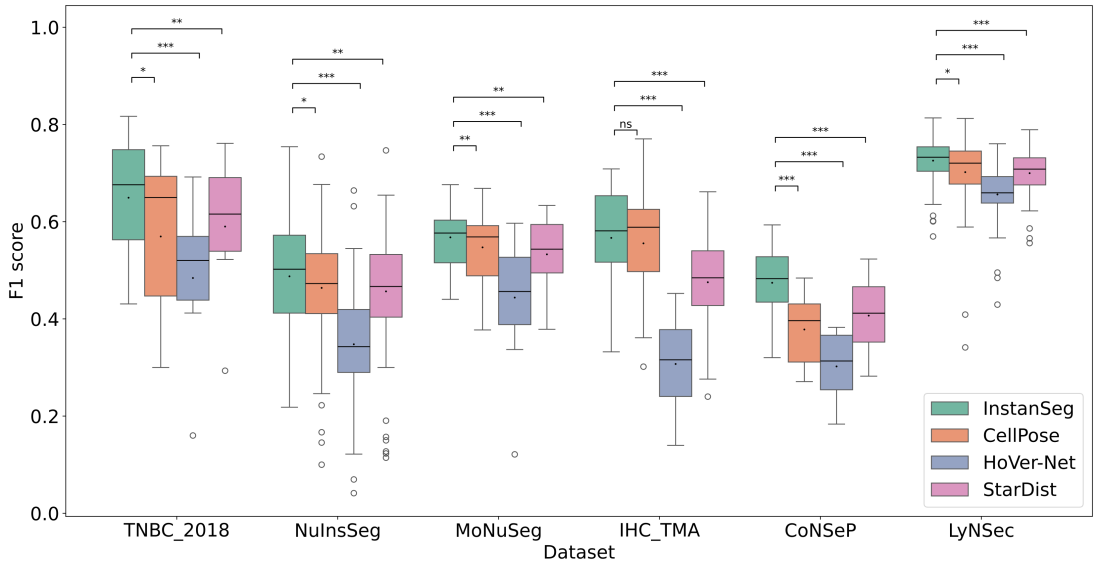

Figure 3.3: Box plot showing the distribution of F_1^{μ} scores on the entire testing set for each of the six datasets. We compare InstanSeg against each of the three other methods using a paired t-test, and report the significance level of the p-value.

Overall, CoNSeP was the most challenging dataset for all the methods as shown by the lowest F1 score. This is likely due to the abundance of crowded, poorly resolved nuclei. InstanSeg performed especially well on this dataset, highlighting our method's robustness in challenging real-world applications. Conversely, LyNSeC was the least challenging dataset, possibly due to the large number of annotations and low variability of images and nuclei shapes, where all methods performed highly. Furthermore, InstanSeg performed well on the small TNBC 2018 dataset, demonstrating its ability to learn from a limited number of high-quality annotations. On most datasets, our method's accuracy can be further improved using test-time augmentations (TTA), although the extra computational costs of applying TTA could be prohibitive in most real-world applications and are not necessary for reaching state-of-the-art accuracy.

We illustrate qualitative segmentation results in Fig. 3.4 and show that detail in the object boundaries is mostly preserved in InstanSeg, while the instance boundaries of Stardist can only be approximated as star-convex polygons. Hence, InstanSeg has the potential to provide additional cellular features of nuclear perimeter and some detail in the granularity of the nuclear envelope as compared to StarDist.

In our comparative analysis, Hover-Net exhibited lower performance across all datasets. This underperformance suggests the method may be less effective as a generalized out-of-the-box solution for nucleus segmentation. Nevertheless, we acknowledge that the model may be improved by fine-tuning training / testing hyperparameters on the individual datasets.

Some segmentation methods, including StarDist (Schmidt et al., 2018), allow for the segmentation of overlapping objects, i.e, assigning single pixels to multiple objects. While, in theory, InstanSeg could allow for the prediction of overlapping objects, we decide to flatten the output of the method by assigning pixels to the most probable of any overlapping instances. This decision was motivated by the fact that current annotated public datasets do not contain any overlapping objects, as well as the lack of computational tools for downstream analysis of overlapping cells in most current software.

Several public segmentation datasets do not report imaging pixel size (e.g. Data Science Bowl 2018 (Caicedo et al., 2019)), and a number of segmentation methods have treated image resolution as an unknown parameter that has to be estimated by the model (Stringer et al., 2021). However, knowledge of accurate pixel size is not only crucial for accurate segmentation and most downstream analysis, it has implications on computational efficiency and memory requirements. Here, we assume that pixel size is recorded during the imaging process, and the image can be resized to a standard pixel size for use with an appropriate InstanSeg model. While this could limit InstanSeg's use for images where the pixel size is unknown, it provides a natural way to apply a model to similar images that have been acquired at different magnifications. It also reduces the risk of false detections arising from structures that may resemble nuclei in appearance but are a completely different size.

Ablations

For investigating the effect of simplifying InstanSeg, we train on all six datasets rather than considering the datasets independently. We report the results in Table 3.3. We show that removing the conditional embeddings ($D_e = 0$) hinders segmentation accuracy, suggesting that the conditional embeddings can capture higher order information relating to instances. We also show that increasing the dimensionality of the positional embeddings beyond the dimension of the image plane ($D_e > 2$) improves segmentation accuracy further.

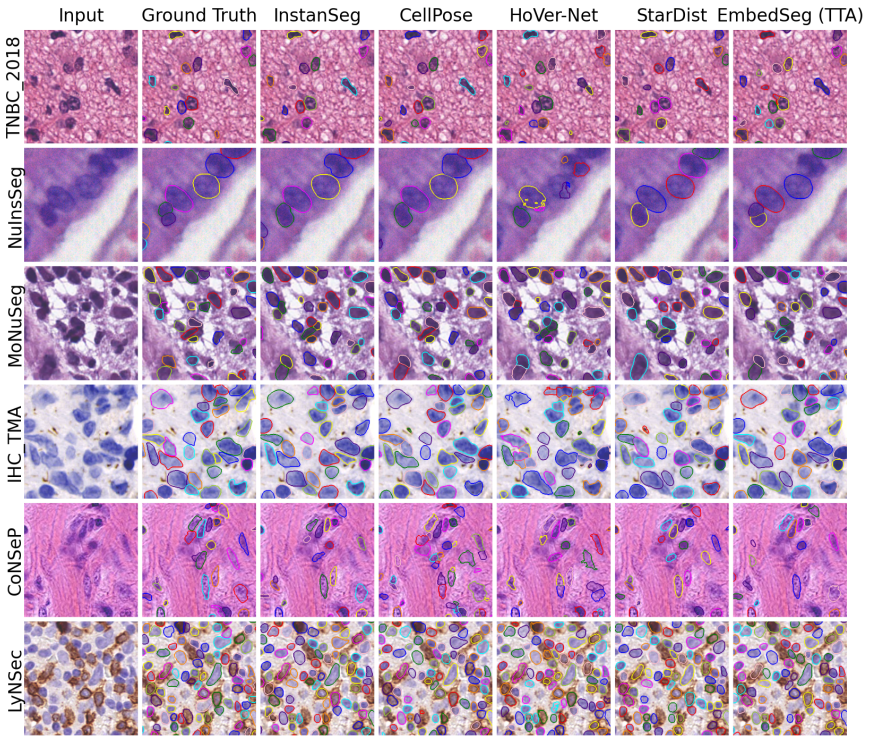

Figure 3.4: Qualitative segmentation results across six nucleus segmentation datasets. For each dataset, each of the four methods was retrained from scratch. The displayed images are 128×128 pixel crops which were randomly selected from the test set.

Table 3.3: Effect of varying the depth of the positional embedding D_e and the conditional embedding D_p. All six datasets were merged for this study.

	F_1^{μ}	$F_1^{0.5}$
$D_e = 4\ D_p = 4$	0.610	0.857
$\hookrightarrow D_e = 4\ D_p = 2$	-0.006	-0.006
$\hookrightarrow D_e = 2\ D_p = 2$	-0.010	-0.011
$\hookrightarrow D_e = 0\ D_p = 2$	-0.020	-0.014

Time efficiency

We profile InstanSeg on the combined test splits of all six datasets, totaling 199 images containing 36,073 instances. Profiling was performed on a laptop GPU (Quadro RTX 3000, 6GB), using mixed precision (FP16) and a fixed batch size of one. We report profiling results in Fig. 3.5. We compare InstanSeg's timing performance to four of the most widely implemented segmentation methods in Table 3.4. We found that InstanSeg was nearly three times faster than the next fastest method (StarDist) and over ten times faster than CellPose. The high inference speed of InstanSeg is in

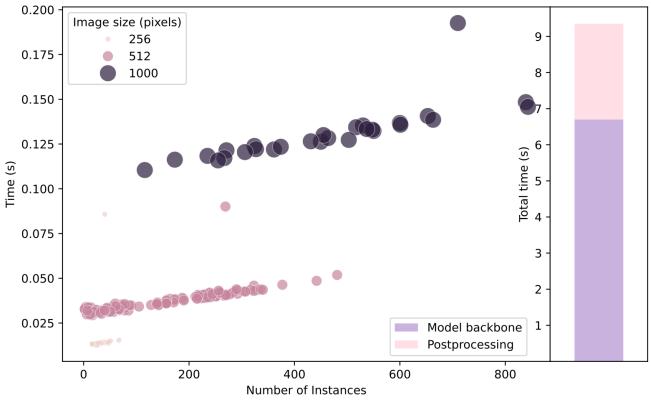

Figure 3.5: InstanSeg inference time is approximately linear with increasing number of detected instances and quadratic with the image dimensions. Timing performed on a laptop GPU. Unlike other methods, the conversion of model outputs to labelled instances is highly efficient, and applying the model backbone (UNet) is the efficiency bottleneck. Timing is for processing 199 images containing 36,073 instances.

large part due to a highly parallelized GPU-accelerated conversion of model outputs to labelled instances. Our postprocessing was over 23 times faster than EmbedSeg's iterative pixel clustering method. Furthermore, InstanSeg only required 1.5 GB of GPU memory, making it suitable for deployment on a wider range of hardware. The efficiency of InstanSeg enables the method to be used on large Whole Slide Images (WSIs). Using the QuPath implementation, inference on a 46,000 by 33,000 pixel image containing approximately 240,000 nuclei took approximately 100 seconds to process using the laptop GPU and similar timing on a MacBook M1 chip. Our Python implementation, which required writing and reading dense segmentation results to disk, took 160 seconds using the Quadro RTX 3000 GPU.

Table 3.4: Time (in seconds) for processing 199 images containing 36,073 instances. Image sizes varied from 256×256 to 1000×1000 pixels. HoVer-Net model time is inflated as the method depends on tiling with large tile overlaps. All the methods were run on a laptop with a Quadro RTX 3000 GPU. StarDist is implemented in Tensorflow, whereas all the others are implemented in Pytorch.

	Model (s)	Postprocessing (s)	Total (s)
StarDist	20.1	6.5	26.6
HoVer-Net	357.6	80.6	438.2
CellPose	-	-	103.2
EmbedSeg	9.4	63.3	72.7
InstanSeg	**6.7**	**2.7**	**9.4**

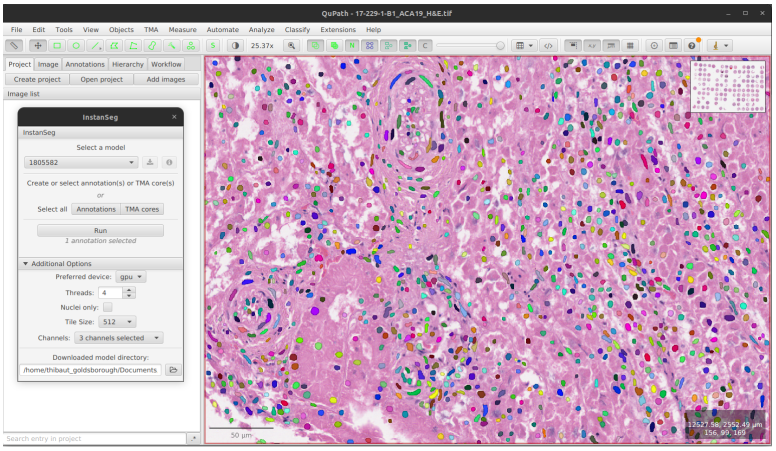

Figure 3.6: Screenshot showing the results of *InstanSeg* extension in the QuPath software. QuPath enables InstanSeg to be integrated into full analysis pipelines with no coding experience required. The extension supports GPU acceleration on both NVIDIA and Apple hardware.

3.4.4 QuPath extension

Unlike most other segmentation algorithms, InstanSeg can be compiled end-to-end using TorchScript, enabling the method to be run outside of the Python environment. We built a QuPath extension, providing a user-friendly interface for running InstanSeg. This can greatly enhance the accessibility of the method to biologists, and enables *InstanSeg* to be easily integrated in full analysis pipelines with limited coding experience. Our extension [2] supports GPU acceleration on both NVIDIA and Apple hardware (Fig. 3.6).

3.4.5 Code availability

The InstanSeg code used to generate the results in this report can be found at tag version 0.1.0[3] of our GitHub repository.

2. https://github.com/instanseg/qupath-extension-instanseg
3. https://github.com/instanseg/instanseg/releases/tag/v0.1.0

3.5 Conclusion

We have proposed a novel method for the segmentation of cells in microscopy images and introduced the use of a neural network to cluster embeddings around optimally selected seed pixels. Our methodology provides several major improvements over previous state-of-the-art embedding-based instance segmentation methods, including higher accuracy, speed and portability. Higher segmentation accuracies will allow for better quantification of cell properties and other downstream analyses, such as classifying cells and interrogating their spatial arrangement in tissue. Furthermore, the improved efficiency of InstanSeg enables the high-throughput evaluation of large image collections on widely available hardware.

InstanSeg's improved portability enables its integration into end-to-end analysis pipelines through user-friendly and open-source software. Unlike most deep learning-based algorithms in the field, InstanSeg is not restricted to running through Python with GPU acceleration using CUDA; rather, it can also be used from other programming languages and includes GPU support on Apple Silicon. By demonstrating the benefits of packaging model inference and postprocessing steps in TorchScript, we hope that our example will encourage more computer vision researchers to develop portable methods that can be readily integrated into other software. Such efforts are crucial to enable standardization and wider adoption by domain experts.

While InstanSeg was designed and optimized for the segmentation of cells and nuclei, its approach can be applied to other instance segmentation tasks, both within and outside the biomedical domain. Relaxed assumptions on the shape of instances may make InstanSeg a good choice for segmenting complex, non-convex structures, while its high efficiency offer benefits for timelapse and video data.

In the next chapter, we extend InstanSeg for the joint segmentation of nuclei and cells across a wider range of wider range of imaging modalities, including highly-multiplexed images with arbitrary number and ordering of imaged biomarkers.

Chapter 4

A novel channel invariant architecture for the joint segmentation of cells and nuclei in multiplexed images using InstanSeg

4.1 Abstract

The quantitative analysis of bioimaging data increasingly depends on the accurate segmentation of cells and nuclei, a significant challenge for the analysis of high-plex imaging data. Current deep learning-based approaches to segment cells in multiplexed images require reducing the input to a small and fixed number of input channels, discarding imaging information in the process. We present ChannelNet, a novel deep learning architecture for generating three-channel representations of multiplexed images irrespective of the number or ordering of imaged biomarkers. When combined with InstanSeg, ChannelNet sets a new benchmark for the segmentation of cells and nuclei on public multiplexed imaging datasets. We provide an open implementation of our method and integrate it in open source software. Our code and models are available on https://github.com/instanseg/instanseg

4.2 Introduction

Multiplexed imaging techniques enable the capture of diverse biological markers, offering unprecedented insight into the protein distribution and cellular composition of tissue. Recently developed high-plex imaging methods such as CODEX (Goltsev et al., 2018), CyCIF (J.-R. Lin et al., 2018) and MIBI (Angelo et al., 2014) enable the capture of dozens of biomarkers in separate imaging channels. While fundamental for the spatial study of tissue, the study of multiplexed images has been hindered by a number of computational challenges. One of these is cell segmentation, in which pixels are assigned to individual cells or cell compartments. Accurate segmentation is crucial for a range of downstream tasks, including cell (Amitay et al., 2023) (Shaban et al., 2024) and tissue (Keren et al., 2018) phenotyping. Consequently, inaccuracies at the initial stage of segmentation can have far-reaching repercussions in subsequent analyses (Qiu et al., 2020).

Despite the pressing need for accurate and generalized cell segmentation methods, the number of computational solutions currently available to biologists is limited. Recent studies applying cell segmentation to multiplexed images have often relied on CellPose (Stringer et al., 2021) and Mesmer (Greenwald et al., 2022). While both deep learning-based models were trained on diverse datasets and reach human-level performance on selected test sets, the segmentation of cells in multiplexed images is typically restricted to using one or two imaging channels (R. Wang et al., 2024). Indeed, Mesmer only accepts a two-channel image, containing one nuclear (e.g DAPI) and one cytoplasmic or membranous marker (e.g E-cadherin). Because it is not possible to rely upon a single marker being available to clearly depict the membrane across all cell types, studies often merge multiple markers into a single channel (Shaban et al., 2024), (C. C. Liu et al., 2023), (Windhager et al., 2023), (Xiao et al., 2021) (Dayao, Brusko, Wasserfall, & Bar-Joseph, 2022) or select a single marker targeting specific cell subpopulations (e.g. CD45) (Jiang et al., 2022). Both approaches inevitably discard information that might otherwise have been informative for identifying cell boundaries. Alternatively, both CellPose and Mesmer can be retrained from scratch with a larger number of input channels. However, such models are less general, and would need to be retrained to match the specific combination of markers used for any image set.

Other methods have approximated cell segmentation masks by expanding nuclear masks obtained from single-channel images (Vázquez-García et al., 2022), (Bankhead et al., 2017). The histoCAT platform (Schapiro et al., 2017) proposes a two-step approach consisting of classifying pixels into three classes (nucleus, membrane and background) using Ilastik (Berg et al., 2019) followed by segmentation using CellProfiler (Jones et al., 2008). This approach requires manually retraining a pixel classifier for images with different biomarker compositions, introducing user-to-user variability. Despite the manual intervention, Ilastik's segmentation based on conventional machine learning has been shown to produce lower segmentation scores on highly variable datasets when compared to deep learning methods (Greenwald et al., 2022).

There therefore remains a need for a generalised computational method that can accurately segment nuclei and whole-cells in multiplexed images. Here, we introduce ChannelNet, a novel channel invariant deep learning architecture capable of generating a fixed three-channel representation of multiplexed images irrespective of the nature, number and ordering of biomarkers. We merge ChannelNet into the InstanSeg method, substantially outperforming previous methods on public datasets in both accuracy and efficiency. We integrate the method within the popular open-source software QuPath (Bankhead et al., 2017) to make InstanSeg (+ChannelNet) amenable for use within existing analysis pipelines, and accessible to biologists with no coding experience.

4.3 Methods

Our contributions to cell segmentation methods are in two parts. Firstly, we build a network that takes an arbitrary number of input channels and generates a fixed three-channel representation. Secondly, we extend our existing InstanSeg method to predict both nuclei and whole-cell labels simultaneously. We combine both methods, such that the fixed three-channel representation serves as input to the modified InstanSeg to obtain a channel invariant cell and nucleus segmentation algorithm.

4.3.1 ChannelNet: a Channel Invariant Network optimized for the analysis of multiplexed images

Multiplexed imaging techniques capture a number of biological markers in separate imaging channels. The number and ordering of these markers varies greatly across datasets, and represents a challenge to convolution-based deep learning methods, which are currently rigid to the number and ordering of input channels. We define *channel invariant* to mean any method that is not fixed to take a specific number of input channels and is invariant to permutations of imaging channels, without discarding potentially useful information relating to nucleus or cell boundaries. We present a novel channel invariant module, which we name ChannelNet, that can be introduced in front of a deep learning backbone and imparts channel invariant properties to the network.

Intuition

We hypothesise that the task of cell segmentation would be straightforward if the input was a noise-free three channel image, where one channel would depict cell nuclei, a second channel depicting the cytoplasm and a third displaying cell membranes as clear separating lines. In practice, each channel in a multiplexed histology image typically represents varying amounts of information relating to 1) cell nuclei specifically (e.g. DAPI), 2) a biomarker localized within one or more cell compartments (nucleus, cytoplasm, membrane) for specific subpopulations of cells only, and/or 3) extracellular structures. In principle, any channel might be informative for cell segmentation, but in a way that varies across cell populations. We do not know in general to what extent any individual channel will be informative for the identification of any specific cell compartment.

To output an informative three-channel image from an arbitrary set of imaging channels, a capable channel invariant network needs to (1) estimate the subcellular location of individual biomarkers, (2) infer the location of cell boundaries, even when membrane markers are not present for all cell populations and (3) suppress uninformative or noisy channels, particularly those that redundantly convey information that is more clearly represented in other channels. While some of these tasks may be achieved by treating channels in isolation, an optimal network would require information from all channels to accurately process each individual channel.

Our Method

We base our channel invariant network on the U-Net (Ronneberger et al., 2015) architecture, popular for its ability to extract high-level features while preserving high spatial resolution. Inspired by earlier work to design deep learning networks that operate on sets (Zaheer et al., 2017), we adopt a novel approach by representing imaging channels as an unordered set of one-channel images, thereby achieving channel invariance by construction. For each block in the U-Net architecture, we first treat the channels in isolation to obtain a number of imaging features, followed by an aggregation step where these features are pooled across all channels and redistributed to the subsequent block by concatenation. The full network is illustrated in Fig. 4.1.

In mathematical notation, if we denote x_d the imaging features corresponding to the d'th channel, then the t'th downsampling block of the ChannelNet architecture performs

$$\begin{aligned} \mathbf{x}_d^{t\prime} &= \mathcal{D}^t(\mathbf{x}_d^{t-1}) \\ \mathbf{x}^{t*} &= \max_d \mathbf{x}_d^{t\prime} \\ \mathbf{x}_d^{t} &= \mathcal{F}^t(\mathbf{x}_d^{t\prime}, \mathbf{x}^{t*}) \end{aligned} \tag{4.1}$$

where $\mathcal{D}$ is a downsampling block with maxpooling and residual connections, and $\mathcal{F}$ is a single convolutional block. The t'th upsampling block performs:

$$\begin{aligned} \mathbf{z}_d^{t\prime} &= \mathcal{U}^t(\mathbf{z}_d^{t-1}, \mathbf{x}_d^{t-1}) \\ \mathbf{z}^{t*} &= \max_d \mathbf{z}_d^{t\prime} \\ \mathbf{z}_d^{t} &= \mathcal{F}^t(\mathbf{z}_d^{t\prime}, \mathbf{z}^{t*}) \end{aligned} \quad , \tag{4.2}$$

where $\mathcal{U}$ is an upsampling block with nearest interpolation and residual connections and $\mathcal{F}$ is a single convolutional block. Our downsampling and upsampling blocks are inspired from the CellPose backbone architecture blocks (Stringer et al., 2021), also used in the InstanSeg backbone architecture.

The network relies on a permutation invariant aggregation operation to pool information across imaging channels. While a number of operations can achieve this, including sum, mean and self-attention, we choose max-pooling based on simplicity and early empirical results, although we expect other operations to yield similar downstream segmentation accuracy.

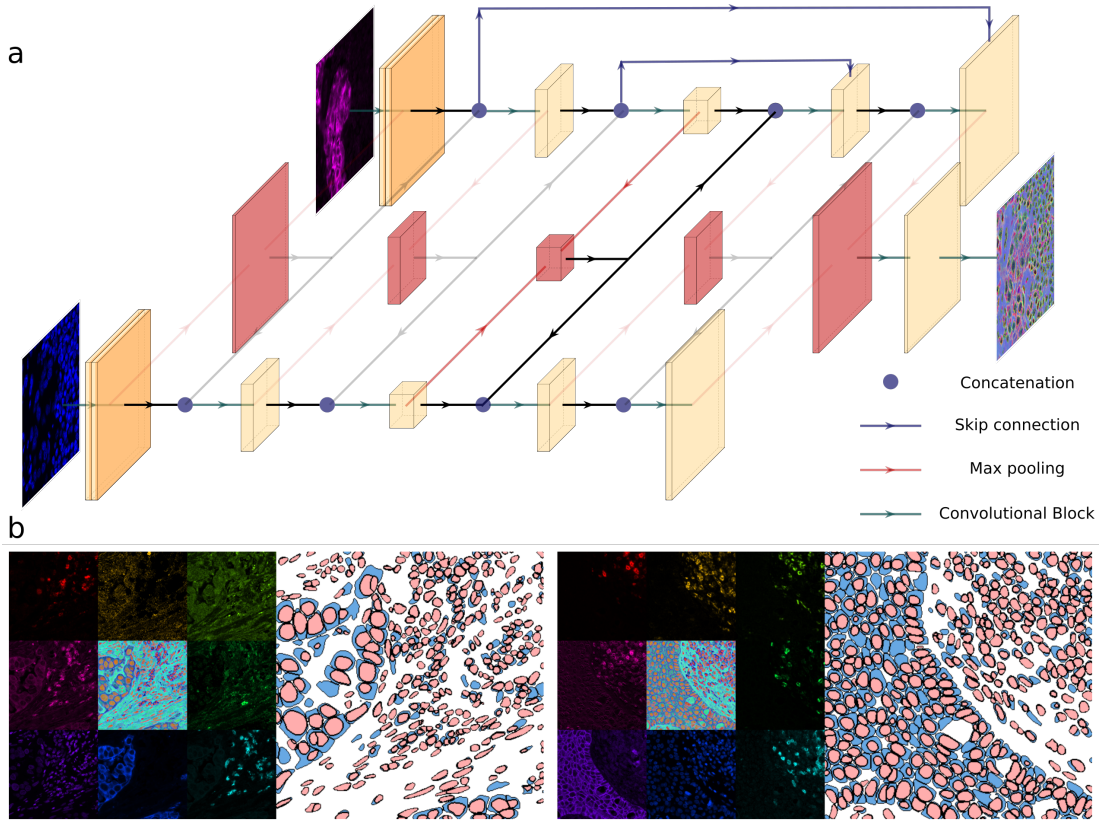

Figure 4.1: (**a**) The ChannelNet network, based on the U-Net (Ronneberger et al., 2015) architecture, achieves channel invariance by construction. A multiplexed image is treated by the network as an unordered set of one-channel images (only two input channels depicted for simplicity). Identical blocks are used to process each channel in isolation to obtain a number of imaging features (yellow blocks), these are then pooled using a max-pooling operation to obtain shared features (red blocks), that are redistributed to each member of the set via concatenation (blue circles). At the final block, all the imaging features are pooled and a final convolutional block is applied to obtain a fixed three-channel output, which serves as input to the main InstanSeg method for cell and nucleus segmentation. (**b**) Two examples from the CPDMI 2023 dataset. For each, the eight channels are shown in isolation, as well as the three channel ChannelNet output (centre), accompanied by the final nucleus and whole-cell segmentation masks predicted by InstanSeg.

4.3.2 Nucleus and whole-cell segmentation using InstanSeg

We extend InstanSeg, our embedding-based segmentation algorithm, to predict both nuclei and whole-cell labels from multiplexed images. To this end, we duplicate the InstanSeg U-Net (Ronneberger et al., 2015) decoder, so that one branch predicts nucleus labels and the second branch predicts whole-cell labels. During training, we only compute and backpropagate the loss using whichever labels are present in the ground truth. This method allows for the simultaneous prediction of nucleus and cell labels even when paired labels are not present in the ground truth.

We treat the ChannelNet network as part of the InstanSeg method, with no additional training steps or loss function required. In effect, ChannelNet is trained to minimize the overall segmentation loss by optimizing the input of the main InstanSeg network. As a result, the three-channel output is not forced to correspond to nuclear, cytoplasmic or membranous signals: any informative representation may be learned, but we expect these to correlate with cell compartments. The number of channels produced by ChannelNet is not inherently limited to three, in theory a single-channel representation suffices to capture the boundaries of the cells and nuclei in a field of view. However, the use of three channels enables the model to capture richer features for each cellular compartments and also benefits from easy visualisation on computer screens.

4.3.3 Datasets

TissueNet (Modified Apache, Non-Commercial) (Greenwald et al., 2022) We use TissueNet v1.1 to benchmark the segmentation performance of InstanSeg. TissueNet is a large dataset collated from a range of microscopy platforms. The images consist of two channels, one containing a nuclear marker (e.g. DAPI) and a separate containing a single cytoplasmic or membrane marker (e.g. E-cadherin).

CPDMI 2023 (CC BY 4.0) (Aleynick et al., 2023) The Cross-Platform Dataset of Multiplex fluorescent cellular object Image annotations (CPDMI) is a dataset of multiplexed fluorescence images from various human organs including lung, breast, pancreas, colon, lymph node, ovary, skin, tongue, sacrum, lymph node, hypopharynx, spleen and tonsil. The images were obtained using the Akoya Vectra, Zeiss Axioscan and Akoya CODEX platforms. Whole cell and/or nucleus annotations are provided for small crops of each of the images based on hand-drawn annotations from multiple annotators and reviewed by a pathologist. Altogether, the dataset contains nearly 50

different biomarkers, with individual images containing between 8 and 32 channels. We use this dataset extensively to develop and validate our channel invariant methods. We combine the Vectra and Zeiss images to form our training and validation splits and reserve all the CODEX images to form the test split.

Table 4.1: Total number of images, nucleus and cell annotation counts for each dataset (Tr: Train,V: Validation,T: Test, N: Nuclei, C: Cells)

Dataset	Channels	Images (Tr/V/T)	Instances (Tr/V/T)
CPDMI 2023 Aleynick et al. (2023)	8 - 32	98 / 25 / 10	15,374 / 4,100 / 1,355 (N) 47,162 / 12,139 / 5,880 (C)
TissueNet Greenwald et al. (2022)	2	2,580 / 3,118 / 1,324	932,591 / 275,495 / 135,633 (N) 988,150 / 294,347 / 145,222 (C)

4.3.4 Benchmarks

The number and order of channels differ throughout the CPDMI 2023 dataset. While this is supported by our proposed approach, no previous cell segmentation methods provide channel invariance to enable direct comparison. Consequently, we benchmark our method in two steps: (1) we compare InstanSeg to existing cell and nucleus segmentation methods on an existing fluorescence imaging dataset with fixed number of channels, and (2) we compare our channel aggregation method to previous channel aggregation strategies on an existing fluorescence imaging dataset with a variable number of channels.

For the first step, we benchmark our proposed ChannelNet + InstanSeg nucleus and cell segmentation pipeline against Mesmer (Greenwald et al., 2022) on the TissueNet dataset (Greenwald et al., 2022), consisting solely of two channel images. We use the public Mesmer model, trained on identical train, validation and test splits. For both Mesmer and InstanSeg, we resize images to 0.5 microns per pixel, as required by the models. Timing was performed on a laptop with a 6GB Quadro RTX 3000 GPU with a batch size of 1.

For the second step, (1) we investigate whether our channel aggregation strategy harms the segmentation accuracy of InstanSeg on the TissueNet dataset, which contains a fixed number of channels. (2) We compare our channel aggregation strategy to other methods that are currently being used to merge channels (Shaban et al., 2024), and (3) investigate how increasing the number of input channels affects the final segmentation accuracy. Finally, (4) we investigate the effect of ablating the ChannelNet architecture on the accuracy of InstanSeg's predictions.

4.3.5 Channel aggregation baselines

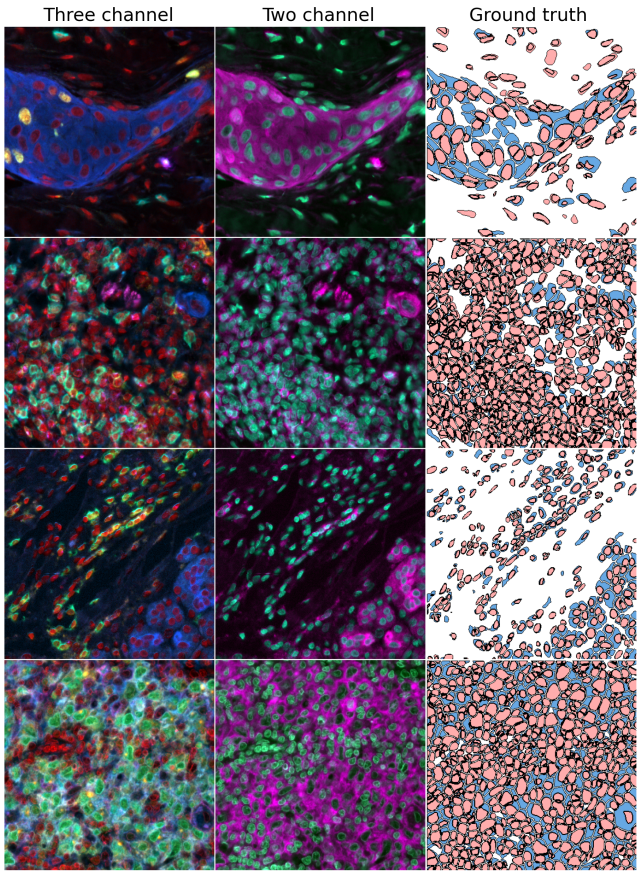

Figure 4.2: Channel aggregation baselines. Three-channel: nuclear biomarkers are kept in a separate channel (here in red) and the other channels are linear projections into RGB space. This is often used in visualisation software. Two-channel: markers are separated based on whether they are mostly expressed in the nucleus or in the cytoplasm, this has previously been used to segment multiplexed images using Mesmer.

Previous studies applying deep learning-based cell segmentation to multiplexed images have generated two-channel images, with one containing a nucleus marker and a second summing the cytoplasmic markers (Shaban et al., 2024) (C. C. Liu et al., 2023). On the CPDMI 2023 dataset, we use the ground truth labels to determine the subcellular location of each marker, and use this information to merge markers ex-

pressed predominantly in the cell nucleus into one-channel and cytoplasmic markers in a separate channel. Specifically, if the mean channel intensity under the nucleus labels was larger than the mean intensity under the whole-cell labels, the marker was determined to be expressed in the nucleus and conversely for cytoplasmic markers. Note that we had to look at the test set labels for determining the location of some biomarkers that did not appear in the training split.

We add a third baseline where we generate three-channel (RGB) images by randomly assigning primary and secondary colours to each of the input channels. This RGB projection is often used for visualization purposes and may enable the separation of touching cells visually. In this baseline, we ensure that the channel corresponding to the nuclear marker is always in the same channel. Examples of the three-channel and two-channel channel aggregation baselines are shown in Fig. 4.2.

4.3.6 Ablation

For our ablation study, we prevent ChannelNet from sharing information across the channels. In practice we set $\mathbf{x}^{t*}$ in Eqn. Eq. (4.1) and $\mathbf{z}^{t*}$ in Eqn. Eq. (4.2) to zeros. This is equivalent to suppressing all but the last red block in Fig. 4.1.

For all our ablations and channel aggregation baselines, we train separate models from scratch on identical train, validation and test splits.

4.3.7 Evaluation metrics

The F_1 score was used as a metric for detection accuracy. We report both $F_1^{0.5}$, determined at the Intersection over Union (IoU) threshold of $\tau = 0.5$ and F_1^{μ}, calculated as the mean F_1 score over the interval $[0.5, 0.9]$ with a step of 0.1. We also report Segmentation Quality (*SQ*) defined as the average IoU of all correct matches (above the IoU threshold of $\tau = 0.5$). Our metrics are calculated using the Stardist implementation[1].

1. https://github.com/stardist/stardist

4.3.8 Preprocessing

We scale the input image so as to set the 0.1% and 99.9% percentiles of the pixel values to 0 and 1 respectively, independently across channels. We resize images to 0.5 microns per pixel using bilinear interpolation.

4.3.9 Augmentations

For the benchmarks on the TissueNet dataset, we use minimal augmentations during training so as to fairly compare results with other methods. These include horizontal and vertical flips, axis-aligned rotations and random crops of size 256×256.

For additional models that we make publicly available, we use further augmentations, including concatenating up to 30 duplicated channels with various amounts of Poisson noise, followed by channel suppression with probability 0.3. We also perform histogram normalization, contrast and brightness shifts.

4.3.10 Training details

We train ChannelNet and InstanSeg in a single training loop comprising 500 epochs, each consisting of 1000 batches of size 3. To batch images with a variable number of channels, we concatenate extra empty channels to make all batched images the same size. We train on 256×256 pixel crops, using the Adam optimizer and a fixed learning rate of 0.001.

4.3.11 Assigning nuclei to cells

Assigning individual nuclei to their respective cells is a non-trivial and often ignored issue in cell segmentation. The difficulty arises when (1) multiple nuclei are predicted within a single cell, (2) a nucleus does not overlap with any predicted whole cell, (3) a nucleus is not fully contained by a corresponding cell or overlaps with multiple cells. As these labeling inconsistencies are present in the training and testing datasets, we do not post-process the segmentation predictions to match individual nuclei and cells. However, we provide the tools to resolve such inconsistencies in our public implementation to facilitate downstream tasks. Based on the observation that nucleus segmentation is typically an easier task compared to whole-cell segmentation, our approach, illustrated in Fig. 4.3, is in two steps. For each cell overlapping with one or

more nuclei, we match it to the nucleus with the highest overlap. We then define the cell mask as the union of the initial cell mask and the matched nucleus mask. For any nucleus that remains unmatched, we create a new cell mask that exactly corresponds to the nucleus mask.

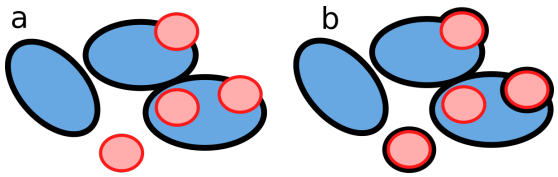

Figure 4.3: Sketch showing the result of our nuclei to cell assignment steps. Nuclei and cells are predicted independently by InstanSeg (a), and can later be resolved for downstream analysis (b). Black lines: cell boundaries, red lines: nucleus boundaries.

4.4 Results

4.4.1 InstanSeg (+ChannelNet) sets a new state of the art for the segmentation of cells and nuclei in multiplexed images

We report our benchmarking results on the TissueNet test set in Table 4.2. We find that the InstanSeg base model, here accepting only two input channels, outperforms Mesmer for both nucleus and whole-cell segmentation. We further show that adding a ChannelNet adaptor, which loses information on the ordering or number of input channels, results in little degradation in the segmentation accuracy.

We also show that InstanSeg is a highly efficient cell segmentation algorithm. The processing speed on the TissueNet test set containing 1,324 images of shape 256 $\times$ 256 pixels was up to 42.7 images per second on a laptop GPU.

4.4.2 InstanSeg (+ChannelNet) allows for the accurate segmentation of cells and nuclei in images with varying number and ordering of channels

We report the segmentation accuracy of InstanSeg on the CPDMI 2023 dataset, comprising images with between 8 and 32 channels, in Table 4.3. InstanSeg predicted nucleus and whole-cell labels with an accuracy of $F_1^{0.5} = 0.818$ and $F_1^{0.5} = 0.752$ respectively. The segmentation accuracy was substantially higher than the two-channel

Table 4.2: Quantitative segmentation results on the TissueNet test set. Note that some of the test set labels were curated using a Mesmer model, as described in Greenwald et al. (2022).

Method	Target	F_1^{μ}	$F_1^{0.5}$	*SQ*	Time (s)	Images/second
Mesmer	Nuclei	0.7115	0.9030	0.8421	280	4.7
	Cells	0.6328	0.8593	0.8163		
InstanSeg	Nuclei	**0.7760**	0.9160	**0.8738**	**31**	**42.7**
	Cells	**0.6725**	0.8699	**0.8343**		
InstanSeg (+ ChannelNet)	Nuclei	0.7649	**0.9207**	0.8646	36	36.8
	Cells	0.6654	**0.8811**	0.8252		

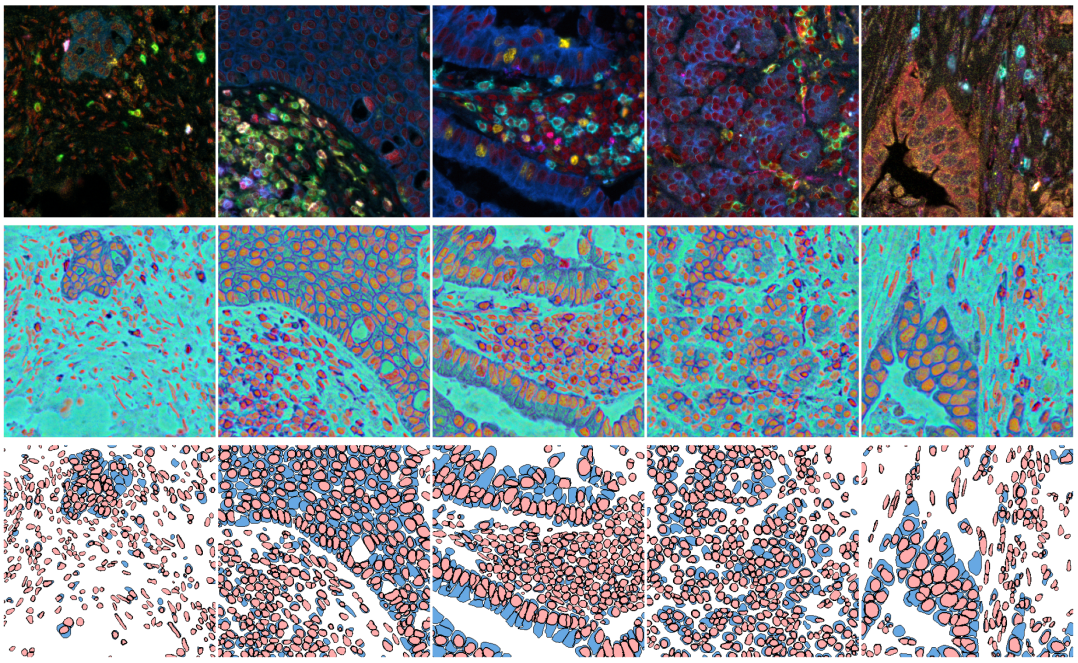

Figure 4.4: Qualitative results of InstanSeg (+ ChannelNet) on the CPDMI 2023 dataset. Top row: five multiplexed images (rendered in RGB for display). Note that the number and ordering or colour channels is not conserved across experiments. Middle row: corresponding fixed three-channel representation generated by ChannelNet. Note the consistency of the nuclear, cytoplasmic and membranous signals. Bottom row: final nucleus and cell segmentation masks predicted by the InstanSeg (+ChannelNet) method.

baseline which separated nuclei and cytoplasmic markers, as used in previous studies (Shaban et al., 2024). Our three-channel-projection baseline performed worse, despite such RGB projections being the standard for visualizing multiplexed images

in software. Compared to the other baselines, our channel aggregation strategy has the advantage of not requiring user input on the subcellular location of the markers. Furthermore, we show that ablating the ChannelNet adaptor degraded segmentation performance, thereby demonstrating the benefit of cross-channel information pooling.

Table 4.3: Baseline and ablation study on the CPDMI 2023 validation set.

Method	Target	F_1^{μ}	$F_1^{0.5}$	*SQ*
InstanSeg (+ ChannelNet)	Nuclei	**0.522**	**0.818**	**0.767**
	Cells	**0.438**	**0.752**	**0.741**
InstanSeg two-channel (No ChannelNet)	Nuclei	0.498	0.798	0.761
	Cells	0.380	0.710	0.716
InstanSeg three-channel (No ChannelNet)	Nuclei	0.330	0.638	0.707
	Cells	0.345	0.662	0.709
InstanSeg (+ ablated ChannelNet)	Nuclei	0.491	0.798	0.757
	Cells	0.418	0.738	0.732

We show qualitative segmentation outputs in Fig. 4.4. As hypothesized, the fixed three-channel ChannelNet outputs correlated with cell compartments, shown by the strong nuclear, cytoplasmic or membranous signals. We stress that ChannelNet is only trained to minimise the segmentation error of the main InstanSeg network, so intermediate representations may differ upon retraining.

InstanSeg (+ChannelNet) segmentation accuracy improves with increasing number of imaging channels

We test InstanSeg (+ChannelNet) on the CPDMI 2023 test set consisting of 28-32 channel CODEX images. We show that the accuracy of the whole-cell predictions increase monotonically as the number of randomly sampled input channels was increased. Our method predicted cellular labels with an $F_1^{0.5} = 0.65$ when a single DAPI channel was provided, and $F_1^{0.5} = 0.73$ when all channels were provided (Fig. 4.5). This result alone is not sufficient to conclude that our method always benefits from increased number of input channels. A similar trend would be expected if ChannelNet was simply selecting the three most informative channels and ignoring the rest, as sampling more channels would inevitably increase the chances of picking the most informative ones. To ensure that ChannelNet actually benefits from more input channels, we first rank the informativeness of each channel independently, and

then sample these in order. We find a similar trend using this method, confirming that ChannelNet does benefit from increased number of input channels. Note that we discarded blank or redundant channels for this study, but found no significant change in accuracy when these were included. We show qualitatively the effect of increasing the number of input channels from one to seven in Fig. 4.6, showing that ChannelNet provides more informative representations when more input channels are included, even for channels that only captured very few cells.

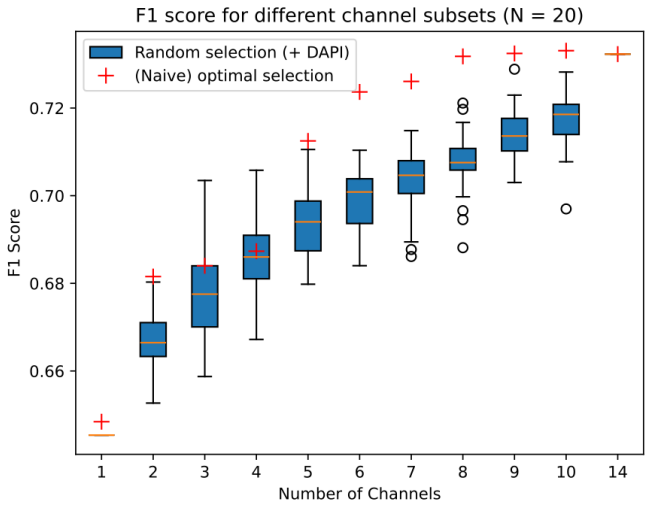

Figure 4.5: On the CPDMI 2023 test set, consisting of 13-14 non-redundant channel images, we show the F1 score of the whole-cell predictions as we increase the number of randomly sampled channels (blue boxes) for N=20 repeats. In each case we ensure that a DAPI channel is selected. Next, we rank each channel by its informativeness based on F1 score when using that channel alone. We then sample the top N most informative channels and evaluate InstanSeg on their combination (red crosses).

InstanSeg enables accurate analysis of multiplexed images.

InstanSeg provides segmentation masks for both nuclei and whole-cells. We find that this dual prediction enables downstream analyses, such as predicting the subcellular location of biomarkers and nucleus to cell area ratios. We compare InstanSeg's predictions against ground truth annotations on the CPDMI 2023 dataset in Fig. 4.7, and show that InstanSeg's downstream predictions strongly agree with those from the ground truth annotations.

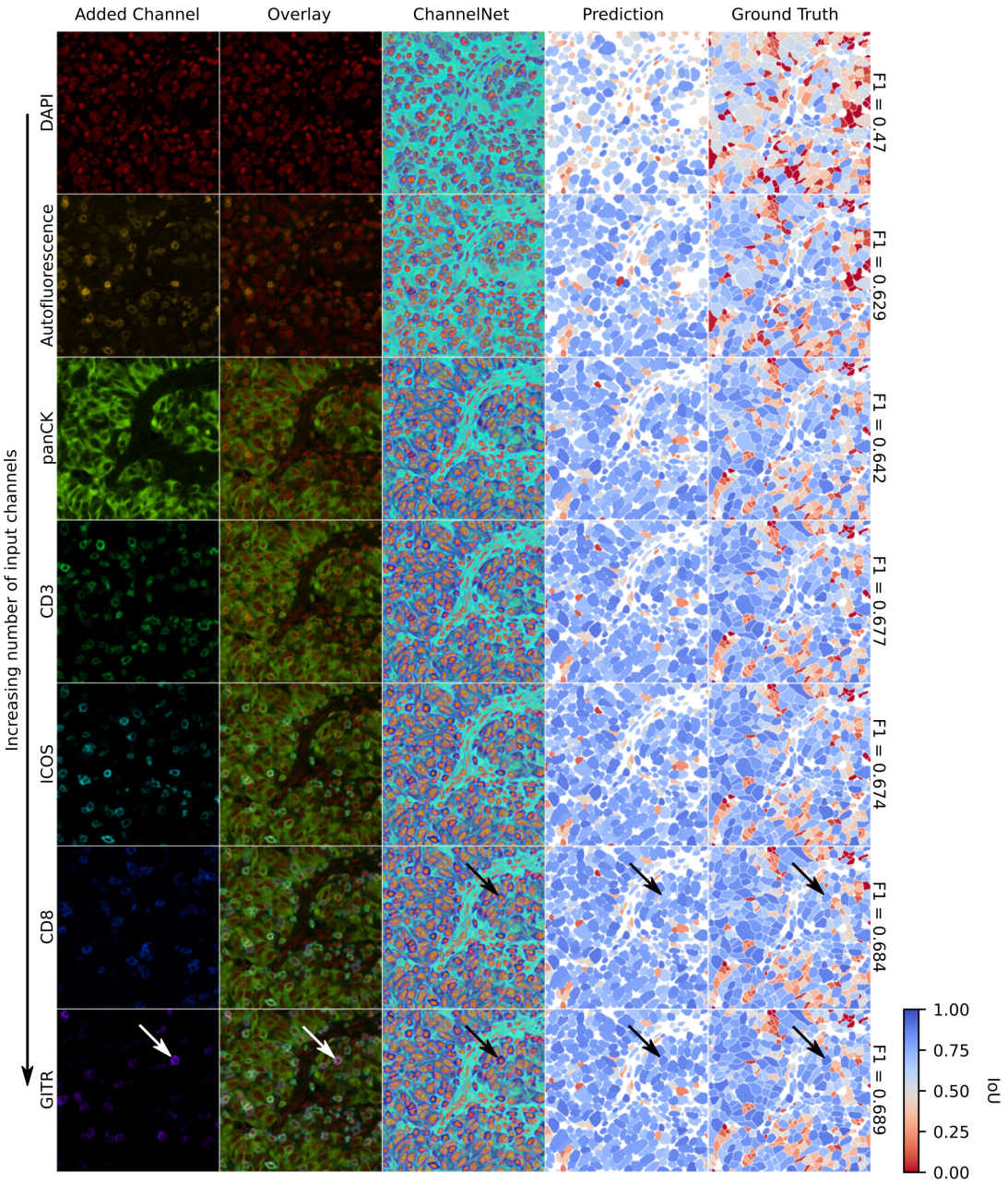

Figure 4.6: Qualitative results showing the effect of increasing the number of input channels from one (top row) to seven (bottom row). The ChannelNet intermediate RGB representations, which serve as input to the main InstanSeg model are depicted in the central column. The last two columns show the predicted and ground truth whole-cell labels, the per-cell agreement is shown using an Intersection over Union (IoU) metric. In other words, instances coloured in red in the columns “Prediction" and “Ground Truth" are false positives and false negatives respectively. Note how markers that were expressed in only some of the cells (e.g. GITR) subtly affected the intermediate RGB representations and eventually allowed for more accurate cellular boundary predictions (see arrows).

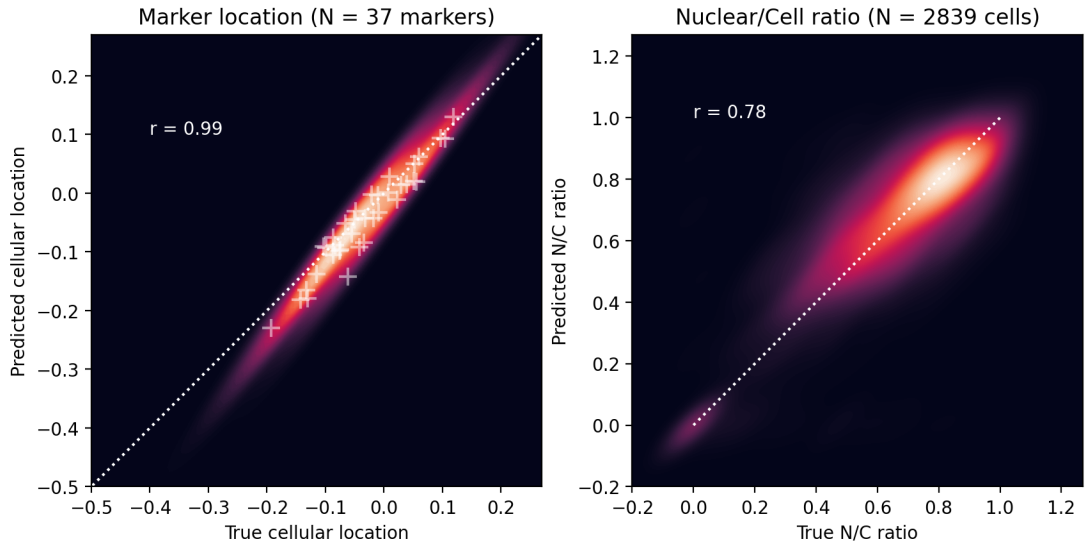

Figure 4.7: Left For each marker in the CPDMI 2023 validation dataset, we compare the predicted cellular location $\log_2(\text{nuclear}/\text{cell})$ of each marker compared to the true location obtained from the ground truth labels. Note that markers/cells were not gated for this analysis. **Right** Predicted nucleus to cell area ratio (N/C ratio) of InstanSeg predictions versus ground truth labels on the CPDMI 2023 validation dataset

4.4.3 QuPath extension

All aspects of the prediction - including pre- and post-processing - are encapsulated in a single TorchScript file, greatly facilitating the implementation of the method into other software. We have demonstrated this by building a QuPath (Bankhead et al., 2017) extension, enabling biologists to use InstanSeg (+ChannelNet) with no coding experience. Our extension supports GPU acceleration on both NVIDIA and Apple hardware, and provides a user friendly interface to select imaging channels for segmentation (Fig. 4.8).

4.5 Discussion

In this work we present a novel method for the simultaneous segmentation of cells and nuclei in multiplexed images. First, we show that InstanSeg improves on the previous state-of-the-art method, Mesmer for both nucleus and cell segmentation on the large TissueNet dataset. Not only was the Mesmer method conceptualized using

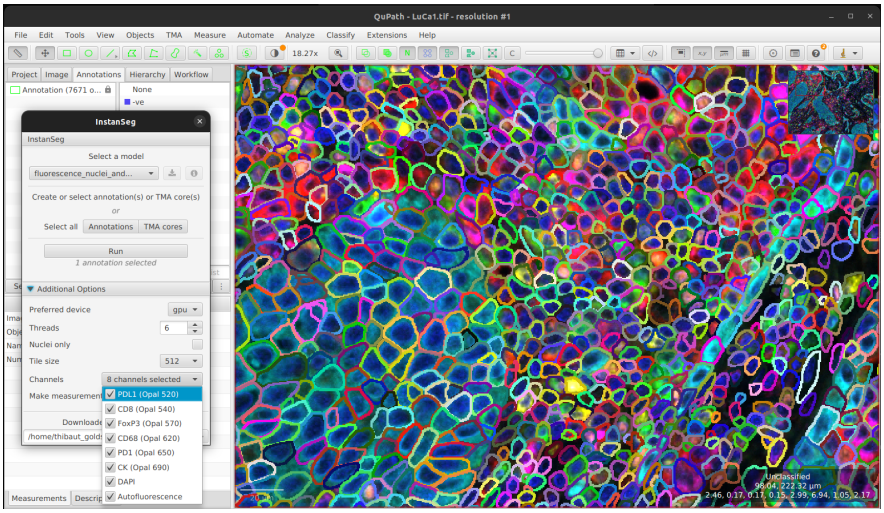

Figure 4.8: Screenshot showing the interactive InstanSeg extension within QuPath. Our extension provides a user-friendly interface for selecting channels for segmentation, and supports GPU acceleration. Sample multiplexed image from LuCa7 ©Perkin Elmer (CC BY 4.0).

this dataset, a number of ground truth annotations in the validation and test sets were obtained using a Mesmer model. There is a possibility that this provided an advantage to Mesmer when benchmarking on this dataset, yet it was still outperformed by InstanSeg.

InstanSeg is a highly efficient segmentation algorithm, which we have previously shown to achieve similar or greater accuracy to CellPose (Stringer et al., 2021), StarDist (Schmidt et al., 2018) and HoVerNet (Graham et al., 2019) for nucleus segmentation, while reducing processing time by at least 60%, as shown in the previous chapter. Here, we retain this efficiency while extending our method to support arbitrary input channels and full cell segmentation. With processing speeds of 42.7 images/second, InstanSeg is nearly ten times faster than Mesmer, which is reported to be among the fastest deep learning-based cell segmentation algorithms (Greenwald et al., 2022). The differences in efficiency are largely due to InstanSeg's lightweight and GPU-accelerated postprocessing step. This eliminated postprocessing as a bottleneck, and contrasts with the strategies employed by CellPose or Mesmer, both of which rely upon a computationally expensive flow tracking or a watershed transform to calculate final outputs.

The proposed combination of ChannelNet + InstanSeg enables the segmentation of cells in multiplexed images irrespective of the number and ordering of input biomarkers. This allows the method to generalize across imaging platforms and experiments. Our channel aggregation strategy not only leads to increased segmentation accuracy, but also eliminates the need for users to manually determine the optimal membrane/cytoplasmic markers for the segmentation step - making it uniquely easy to apply to new image sets. While histoCAT (Schapiro et al., 2017) already enables the prediction of three-channel intermediate representations of multiplexed images, the method requires manual retraining of an Ilastik pixel classifier for different channel combinations. Unlike InstanSeg, this approach risks introducing user-to-user variability and slows down the segmentation workflow. By incorporating additional input channels for segmentation, InstanSeg can more accurately resolve cellular boundaries between biomarkers. We expect that this will allow for improved feature extraction and phenotyping of cells in multiplexed images. Better quantification of cellular properties allows for improved downstream analyses, such as the study of cell interactions within their microenvironments.

While this work focused solely on segmentation, we suspect that the three-channel representations produced by ChannelNet could have uses beyond the identification of cellular boundaries. For example, these representations capture cellular morphologies consistently across biomarker panels which could be used for cell phenotyping. These intermediate representations could also be used to ease visualisation of highly multiplexed images on computer displays.

Openness and accessibility are central to this work. Our development of ChannelNet benefited from the creators of CPDMI 2023 making a large and heterogeneous multiplexed dataset of hand annotated cells and nuclei freely available (Aleynick et al., 2023). Good training and validation data are crucial, and to our knowledge this is the first such dataset to be shared under a permissive open license. By making our own code freely-available and open-source, and by providing a model pre-trained on CPDMI through a user-friendly QuPath extension, we anticipate that ChannelNet + InstanSeg will provide a new standard baseline for multiplexed cell segmentation in the research community. As the method is applied independently to a wider range of images – acquired using different technologies, to look at even more tissues and

markers – we expect that this real-world validation will quickly identify areas where improvement is still required. Our hope is that this will help further accelerate progress by focusing community effort on unsolved problems, and inspire the sharing of new open datasets to train and validate future models.

4.6 The InstanSeg U-Net encoder can be replaced with the SAM encoder

We have demonstrated that InstanSeg achieves state-of-the-art performance for both brightfield nucleus segmentation and for the joint segmentation of nuclei and cells in multiplexed images. We emphasise that the performance improvements of our method are not based on innovations in our U-Net encoder, training optimizations or augmentations, but due to our novel embedding-based training paradigm combined with our novel method for handling unordered multiplexed fluorescence images. We hypothesise that the results we have presented, notably improved accuracy, speed and portability, will generalise to other encoder architectures such as the much larger Segment Anything Model (SAM) (Kirillov et al., 2023), which has recently received considerable attention in the cell segmentation literature (Pachitariu, Rariden, & Stringer, 2025), (Israel et al., 2025), (VandeLoo et al., 2025).

This reflects a general trend in computer vision away from smaller purpose built models towards multi-purpose foundational models. Foundational models typically contain substantially more training weights (SAM ViT-H has ≈ 600 million parameters compared to ≈ 6 million parameters for a typical U-Net.) and often requires collections of multiple training datasets for high generalisation performance. Such models typically require advanced, specialised hardware for training, often relying on remote compute servers, which can limit accessibility for many medical institutions that handle sensitive data. In contrast, smaller, task-specific models can be trained and fine-tuned on more widely available hardware, which facilitates rapid prototyping and local adaptation to new tissue types or imaging modalities. Hence, we view foundational models and smaller purpose-build models as largely complementary, which the former providing better zero-shot performance and generalisation across image modalities and downstream tasks, while the latter provides better efficiency and usability.

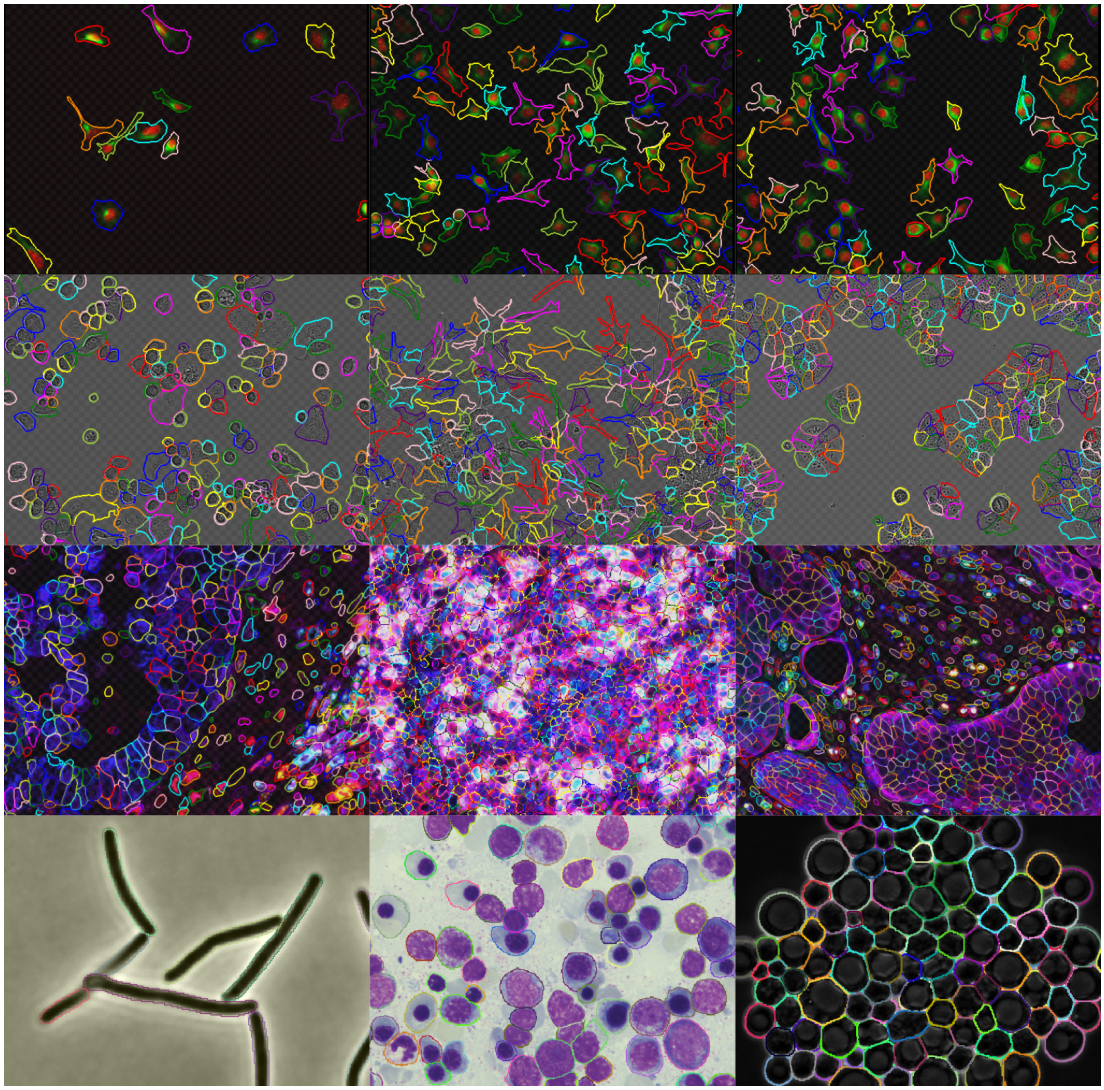

Figure 4.9: Qualitative segmentation results of InstanSeg cell segmentation model with a SAM-based encoder on four public datasets. From top to bottom, by row, Cellpose dataset, LIVEcell, CPDMI 2023 and NeurIPS CellSeg. A single InstanSeg model with a ChannelNet adaptor was used for cell segmentation across a range of imaging modalities, cell and tissue types and experimental conditions.

As a proof-of-concept, we replace the U-Net encoder with a pre-trained SAM encoder (sam_vit_b_01ec64) and fine-tune it using the InstanSeg training objective, using seven cell segmentation datasets: CIL (Yu, Lee, Hariharan, Bu, & Ahmed, 2025), CPDMI 2023 (Aleynick et al., 2023), HPA (Kaimal, Thul, Xu, Ouyang, & Lundberg, 2024), Cellpose (Stringer et al., 2021), LIVEcell (Edlund et al., 2021), Neurips CellSeg (Ma et al., 2024) and Tissuenet (Greenwald et al., 2022). During training, we use a fixed learning rate of 5e-5. While we do not provide quantitative comparisons with CellposeSAM as the authors of CellposeSAM have not yet released explicit testing splits, we present qualitative results that illustrate a pre-trained SAM encoder can provide high generalisation performance with the InstanSeg framework.

In practice, a SAM-based encoder poses several disadvantages compared to U-Net architectures: (1) training and fine-tuning requires advanced GPU hardware, (2) inference is roughly an order of magnitude slower on GPU and even slower on CPU, (2) ViT architectures require a fixed input size (in our case 256×256 pixels), and their lack of translational invariance can introduce tiling artefacts when processing large microscopy images. Furthermore, training across multiple datasets with differing licenses (e.g., CellSeg: NC-ND; TissueNet: custom NC) complicates model sharing, is incompatible with the open-source definition, and puts the end user at risk. Despite these limitations, our experiments demonstrate that InstanSeg can leverage large pre-trained ViT encoders and provides a foundation for future quantitative benchmarking.

In summary, while we anticipate that using InstanSeg with a SAM encoder could be advantageous in certain scenarios, particularly when a single, highly generalised model is needed for microscopy images that are small enough to avoid tiling. However, for large images, especially whole slide images, we expect that the performance gains from using a U-Net encoder with more specialised training will make it the preferred choice.

4.7 Conclusion

In conclusion, we have proposed a nucleus and cell segmentation method for multiplexed images using a novel channel invariant module. Our methodology demonstrated substantial improvements in accuracy and efficiency on public segmentation datasets. The combination of ChannelNet + InstanSeg allows for inference on multiplexed images where the number and ordering of channels need not match

that encountered during training. Our implementation does not require retraining or any manual intervention, thereby reducing inter-user variability and simplifying segmentation workflows. By providing open-source implementations for both Python and QuPath, in addition to pre-trained models, we have provided tools for biologists to efficiently integrate our method in full analysis pipelines.

Chapter 5

The MONKEY challenge: Cell classification in brightfield images using weakly-annotated labels

5.1 Abstract

The segmentation of nuclei and cells is often only the first step in many image analysis pipelines in histopathology. Downstream analyses frequently involve cell classification, querying the spatial arrangement of cellular phenotypes, and ultimately tissue phenotyping for patient diagnosis or predicting therapy response (Vanea et al., 2024). In this chapter, we develop a method for the segmentation and subsequent classification of cells in histopathology images. To facilitate the comparison with other methods and as an opportunity to contribute to pathology research, we enter the public competition: Machine-learning for Optimal detection of iNflammatory cells in the KidnEY (MONKEY), hosted on the Grand Challenge platform[1], running from 20 September 2024 to 9 February 2025. Our solution involved a two-stage segment-then-classify pipeline that used InstanSeg for segmentation and leveraged paired immunohistochemistry and PAS staining for training large convolutional classifiers. Our method achieved the 2nd place for the detection of inflammatory cells (Task 1) and 1st place for the detection of monocytes and lymphocytes (Task 2) in the final competition leaderboard.

1. https://monkey.grand-challenge.org/

5.2 Introduction

Organ transplants can fail because the recipient's inflammatory reaction leads to organ rejection. A key histopathological indicator of rejection is the inflammatory response observed in transplant biopsies (Midden et al., 2024). In the case of kidney transplants, pathologists have developed a classification system for the stratification of inflammatory responses in biopsies called the Banff classification (Haas et al., 2018). The Banff classification is currently the gold standard for diagnosing and grading kidney transplant rejections. The classification scheme focuses on histopathologic evidence of both T cell–mediated rejection (TCMR) and antibody-mediated rejection (ABMR). To evaluate TCMR, the classification system depends on the semi-quantitative assessment of inflammatory cell density and composition in a range of renal compartments. While the correct and efficient assessment of the Banff score is crucial for patient diagnosis and treatment, large inter-institution and inter-pathologist discrepancies have been reported in practice (Furness & Taub, 2001). The aim of this competition was to determine whether computational methods could provide an efficient and consistent alternative to the manual assessment of kidney biopsies.

The MONKEY challenge involved the detection of inflammatory cells in periodic acid-Schiff (PAS)-stained whole-slide images (WSIs) of renal biopsies. While inflammatory cells include polymorphonuclear cells (neutrophils, eosinophils, and basophils) and mononuclear cells (lymphocytes and monocytes), the challenge focused only on the detection and classification of lymphocytes and monocytes. PAS is an inexpensive, standardised and widely available stain, but similarly to H&E, PAS is non-specific and only broadly captures nuclear and membrane structures. As a result, PAS does not immediately allow for the identification of immune cells. For this reason, in addition to each PAS stained slide, a registered immunohistochemistry (IHC) slide is included. The IHC slides include a CD3/CD20 double stain for lymphocytes, resulting in dark brown cytoplasmic staining, and PU.1 stain for monocytes, resulting in a nuclear magenta staining. IHC staining and registration provide a valuable reference for guiding ground truth annotation and algorithm validation. In contrast to PAS, IHC staining can be expensive, time consuming and highly variable, impeding their widespread use in diagnostic settings (O'Hurley et al., 2014). As such, we seek an algorithm that can identify morphological structures in PAS images that are indicative of immune mononuclear cell types.

The task of cell detection and classification can be approached using a range of machine learning methods. In the histopathological setting, methods often involve the open-source software QuPath (Bankhead et al., 2017), which uses traditional image processing methods for the segmentation of cells and extraction of cell features such as area, circularity and stain intensities. Machine learning algorithms such as fully connected neural networks or random forests have then been used for the automated classification of cells based on these precomputed features (Acs et al., 2019). However, accurately classifying the cell types in the MONKEY challenge requires more powerful algorithms that can capture the subtle morphological and staining features that distinguish immune mononuclear cell types. Deep learning-based methods in the bioimage analysis community include Hovernet (Graham et al., 2019) and StarDist (Weigert & Schmidt, 2022), which both treat the task as a semantic segmentation problem by predicting class probabilities at a pixel-level resolution. When combined with instance segmentation, average probabilities for each object can be determined. While these methods have shown high accuracies in the histopathology context, they are difficult to extend to new cell types or staining types without having to retrain both the instance segmentation and classification heads. In general, instance segmentation ground truth is substantially more time-consuming to annotate than point annotations, which limits the size of the training datasets for these methods.

Alternative methods can be used to directly predict coordinates and classifications such as DETR (Carion et al., 2020) or YOLO (Redmon, Divvala, Girshick, & Farhadi, 2016), (Khanam & Hussain, 2024), but it is not trivial to train these networks with weak annotations, such as the paired IHC - PAS images provided in this challenge.

Methods that do leverage paired image modalities often involve virtual staining. For example, S. Liu et al. (2021) introduced the use of Generative Adversarial Networks (GANs) for generating virtual IHC stained images from H&E images. A number of subsequent virtual staining methods were subsequently developed, reviewed in Klöckner et al. (2025). Due to their reliance on GANs or diffusion models (Li et al., 2024), these methods can be difficult to integrate with downstream tasks such as cell classification, as required in the MONKEY Challenge.

We seek an alternative simple training paradigm that (1) decouples instance segmentation from instance classification and (2) can be extended to leverage the large number of paired IHC and PAS images provided in the MONKEY challenge. To this end, we develop cell classifiers that take a small image crop and a binary mask representing a cell of interest and produce a single classification logit for each cell, without requiring any postprocessing steps. We show that this paradigm can be used to leverage paired IHC and PAS images for improved cell classification.

5.3 Methods

Our method was in two parts, first we used an InstanSeg model for the segmentation of all nuclei per PAS-stained slide ROI, which allowed for the extraction of small image patches centred on each nucleus of both the PAS and the co-registered IHC-stained slides. In a second part, we use a deep-learning classifier for the classification of these image patches into three classes: monocytes, lymphocytes and other.

While a classifier could be trained with only manually annotated ground truth, we speculate that an ideal challenge solution would leverage the extensive and rich information provided by the unlabelled but co-registered IHC images to improve cell classification in PAS images. Hence, we seek a method to transfer phenotypic information of nuclei from the IHC-stained images to the PAS-stained images.

A simple method could involve stain deconvolution (Ruifrok, Johnston, et al., 2001) of the IHC images for the automatic classification of CD3/CD20 positive (brown) cells and PU.1 positive (magenta) cells based on the mean stain intensity of each cell. However, due to the huge variability in the staining pattern in the IHC images, we seek a more robust method to determine stain positivity. To this end, we use manually annotated ground truth to train a convolutional model to predict stain positivity in IHC images, and use this model to automatically annotate a large number of cells in the precisely registered PAS images. We then use this weakly annotated dataset to train a separate model for the classification of cells using only PAS images. An overview of our method is shown in Fig. 5.1.

Our method bears similarity to the technique of knowledge distillation (Hinton, Vinyals, & Dean, 2015), in which a large, high-performing model or ensemble of models is used to train a typically smaller student model. A related multi-modality knowledge distillation approach was introduced by Hu et al. (2020), where information from multiple data sources is used to train a student model. In our setting, the high performance of the teacher model comes from the high information content of the IHC modality, rather than from increased model size or complexity.

5.3.1 Dataset

The MONKEY challenge provides both PAS and IHC stained images as well as manually annotated labels for training and evaluating models. Furthermore, the challenge rules allows for freely and publicly available external datasets and pre-trained models to be used in the competition.

Gold standard dataset

The challenge dataset consists of 153 WSIs collected across six different pathology departments, with four departments used for model training and two departments used for testing. Manually curated ground truth cover 231 regions of interest (ROIs) with dot annotations for approximately 30,000 monocytes and 60,000 lymphocytes. For each slide, a paired PAS and IHC WSI is available at a resolution of 0.24 μm pixels, which were previously registered using the HistokatFusion software (Budelmann, Weiss, Heldmann, & Lotz, 2022). Within each ROI, a publicly available InstanSeg model *brightfield_nuclei*[2] was used for the segmentation of all nuclei. This model was trained on the *TNBC 2018* (Jack et al., 2021), *NuInsSeg* (Mahbod et al., 2024), *IHC TMA* (R. Wang et al., 2024), *CoNSeP* (Graham et al., 2019) and *LyNSeC* (N. Hussein et al., 2023) datasets with heavy augmentations. Approximately 150,000 nuclei detected by InstanSeg did not correspond to monocyte or lymphocyte point annotations and were hence classified as *other*.

2. `https://github.com/instanseg/instanseg/releases/download/instanseg_models_v0.1.0/brightfield_nuclei.zip`

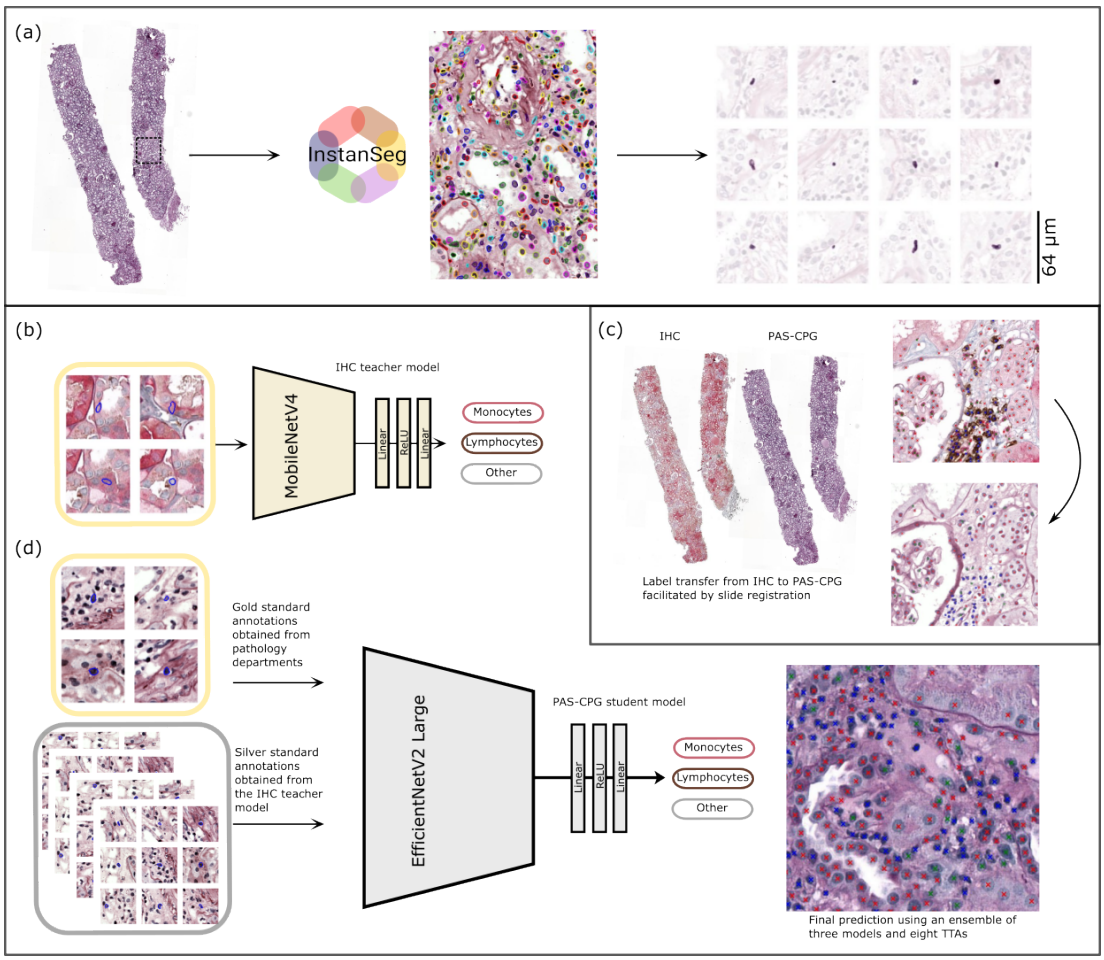

Figure 5.1: Pipeline diagram of our solution to the MONKEY challenge. (a) InstanSeg was used to segment all nuclei within ROIs. For each detected nucleus, a four-channel patch was extracted. (b) A MobileNetV4 model was trained to classify nuclei patches in the IHC images. (c) Precise registration allowed for the transfer of predicted phenotypes from IHC images to PAS images, enabling the creation of a new dataset. (d) We trained a large EfficientNetV2 model on the combined datasets.

Silver standard dataset

We generate a large dataset of weakly annotated image crops using the registered IHC WSI. Specifically, for each WSI in our training set, we load 1000 non-overlapping image square ROIs of side 1024 pixels. We ensure the sampled ROIs contain at least 50% foreground tissue regions using an Otsu threshold (Otsu, 1979) of the WSI thumbnails. We produce an instance segmentation map of detected nuclei by running InstanSeg on the PAS images. For each detected nucleus, we extract a 128 pixel image crop for both the IHC and PAS-stained WSIs. We infer object classes using a trained teacher model and eight test-time augmentations (TTA) on the IHC crops. Due to a large class imbalance and storage limitations, we undersample the majority classes. Specifically, within each ROI, we select an equal number of samples from each predicted class, based on the size of the smallest class plus a small margin (10 additional samples). This ensures a roughly balanced subset without entirely ignoring ROIs that do not contain all three predicted classes. PAS-stained crops and model predictions are then saved in the *.h5* format for efficient downstream loading. In total, our silver standard dataset contained 700,000 monocytes, 900,000 lymphocytes and 1,250,000 *other* cells.

5.3.2 Models

We seek suitable classifiers for task of (1) cell classification of IHC-stained images using the gold standard annotations and (2) cell classification of PAS-stained images using the gold and silver standard annotations. We hypothesise that the first task is relatively easy, as IHC staining is directly indicative of cell type. To avoid overfitting on this task, we seek a model with relatively few parameters. We call this model *teacher model* as it is used to generate the silver standard annotations. We hypothesise that the second task is substantially harder, as cell types in PAS-stained images need to be determined from subtle morphological differences that are often not visible to the human eye. As a result, we seek a much larger model for this task which we call *student model*, and train this model using the much larger silver standard annotations as well as the gold standard annotations.

Teacher model

We empirically compare various architectures available on timm (Wightman, 2020) and found that the larger convolutional models would quickly overfit on the smaller gold standard dataset, we ultimately select the MobileNetV4 "conv large" model containing approximately 31 million trainable parameters. To accommodate for small image patches, we halve the stride of the first convolutional layer from two to one. We do not use pre-trained weights for this model.

Student model

We use the EfficientNetV2 large from torchvision (Paszke, 2019) containing approximately 120 million parameters. This was the largest convolutional model with pretrained weights available in timm or torchvision. We empirically found that larger ViT models failed to converge with the same training hyperparameters used for EfficientNetV2. Similarly to the teacher model, we halve the stride of the first convolution from two to one. We use the pre-trained weights from torchvision on imagenet 1k (Deng et al., 2009).

5.3.3 Training objective

For training of the teacher model, we use standard cross entropy in a three-class classification problem. Hence, for each nucleus patch, the teacher model outputs three logits. When training the larger student model, we use both gold and silver standard datasets. We hypothesise that the *weak* labels from the silver standard dataset cannot be treated interchangeably with the gold standard labels due to (1) deliberate class imbalance during the sampling strategy of the silver standard dataset, (2) potential off-target staining of the IHC stains and (3) possible systemic bias introduced by the teacher model. As a result, we train the student model to predict six logits, three for the gold standard labels and three for the silver standard labels. During training, we mix batches with samples from both datasets, and only backpropagate through whichever label is available. At test time, we only use the three logits corresponding to the gold standard labels.

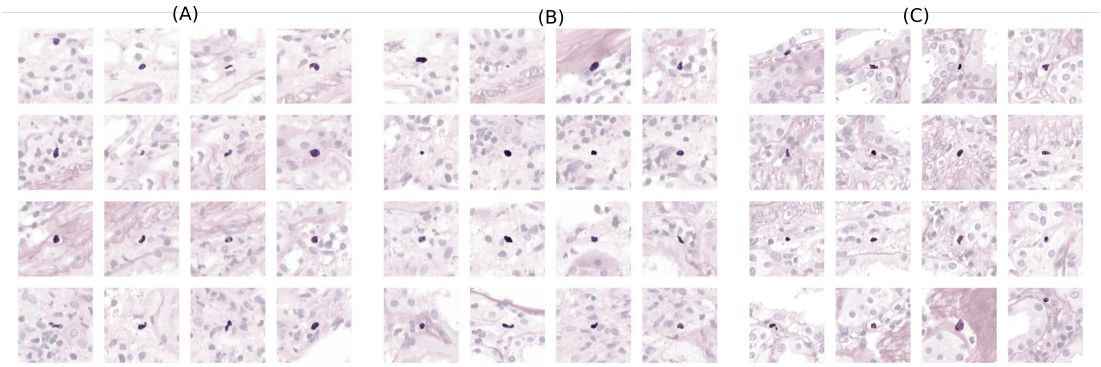

Figure 5.2: Small image crops each containing a cell of interest, illustrating the difficulty of cell classification based on PAS stained images alone. We show 16 monocytes (A) and 16 lymphocytes (B). We reveal the class of the 16 cells in (C) in the footnote, to allow the reader to attempt classifying these cells. Monocytes are stereotypically bean-shaped in appearance, while lymphocytes are typically smaller, denser and rounded nuclei. While the crops used for training have four channels (three RGB channels and one binary mask channel), we display the fourth channel by decreasing the opacity of the pixels outside of the cell mask.

5.3.4 Training parameters

For all the models, we use a batch size of 128, the Adam optimizer (Kingma & Ba, 2017) and a learning rate of $1e-3$ and weight decay value of $1e-4$. We train each model until convergence on our validation set.

5.3.5 Augmentations

We use minimal augmentations in our training pipeline, including flips and rotations. We found that the 128 × 128 pixel patches were not sufficiently large for standard stain normalisation or stain augmentations (Otálora et al., 2022) to be performed on the fly during training. We perform no data preprocessing or manipulation steps other than dividing the 8-bit images by 255. At test time, we use eight TTAs including flips and rotations and aggregate the model outputs by taking a mean.

---

2. The cells in (C) are monocytes

5.3.6 Metrics

For internal validation of classification accuracy, we use the standard F1 score. The final detection metric was the Monai (Cardoso et al., 2023) implementation of the Free Response Operating Characteristic (FROC). The FROC curve plots the false positive (FP) rate with the true positive (TP) rate over the entire range of class probability thresholds. A TP is recorded when a predicted dot coordinate is within a fixed error margin of a ground truth dot coordinate ($5\,\mu\mathrm{m}$ for monocytes, $4\,\mu\mathrm{m}$ for lymphocytes, and $5\,\mu\mathrm{m}$ for inflammatory cells). The FROC curve operates across a range of prediction thresholds which makes it less sensitive to differences in model calibration than a single-threshold metric such as F1 score, making it more suitable for comparison of different methods.

5.3.7 Other winning entries of the MONKEY challenge

TIAKong

The TIAKONG method used an ensemble of encoder-decoder models with EfficientNetV2 backbones. Specifically, the team used three separate decoders for the detection of inflammatory cells, monocytes and lymphocytes. For each of these, the decoders predicted three separate semantic maps corresponding to object centroids, object masks and object contours. To generate the ground truth annotations, the team used a pre-trained Nuclick model (Koohbanani, Jahanifar, Tajadin, & Rajpoot, 2020). No additional information about the method was published at the time of writing.

Aira Matrix

Aira Matrix (Deotale, Ambast, Ramchandani, Das, & Thomas, 2025) used an ensemble of two recent object detection models DEtection TRansformer (DETR) (Carion et al., 2020) with a pretrained Swin-L Backbone (Z. Liu et al., 2021) and YOLOv5-L (Khanam & Hussain, 2024). The authors used binary cross-entropy for object detection, focal loss for classification in DETR, cross-entropy for classification in YOLOv5-L, Smooth L1 loss for bounding box regression in DETR and Generalised Intersection over Union (GIoU) loss for YOLOv5-L. To ensemble the two models, the Weighted boxes fusion (WBF) method (Solovyev, Wang, & Gabruseva, 2021) was used. This method used substantial augmentations including geometric, colour and contrast adjustments.

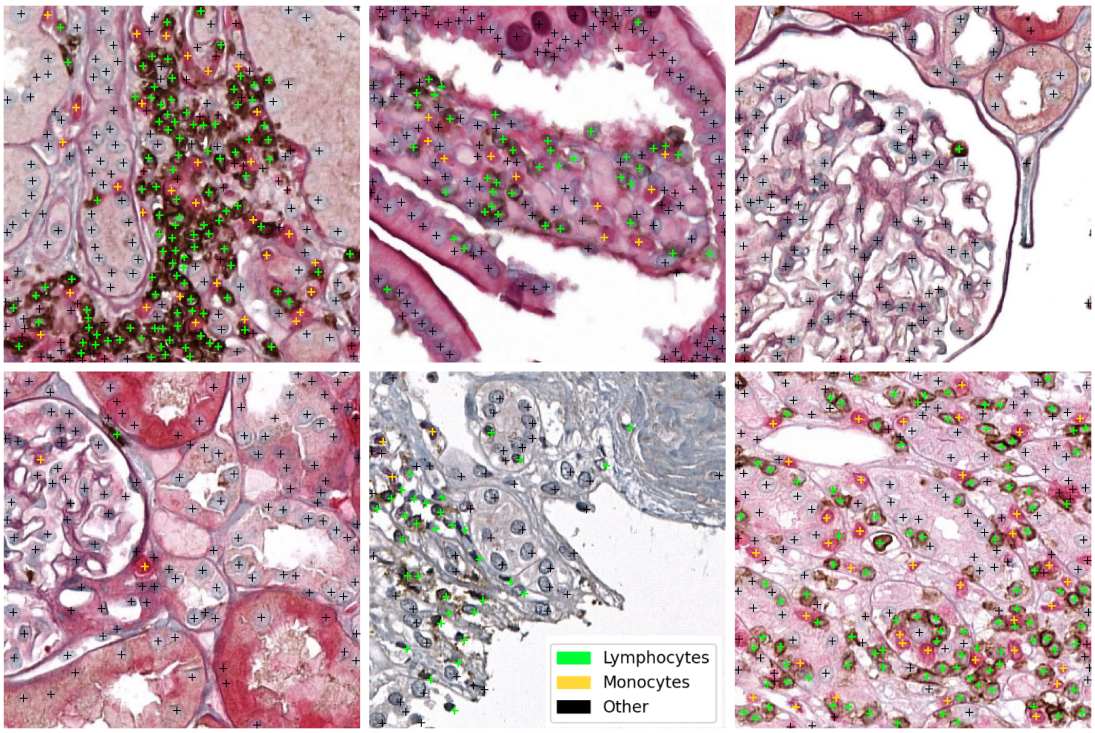

Figure 5.3: Qualitative results of the teacher model on the IHC stained images. The IHC images included a CD3/CD20 double stain for lymphocytes, resulting in dark brown cytoplasmic staining and PU.1 stain for monocytes, resulting in a nuclear magenta staining. Note the huge variability in staining pattern and intensity across the ROIs, which were randomly sampled from different WSIs. This variability, as well as off-target staining (e.g. large magenta nuclei in the top row, centre column) makes it difficult to infer stain positivity based on image intensity values alone. Despite the stain variability, the trained classifier was able to predict cell classes with high accuracy.

5.4 Results

As a baseline, we report classification results of the smaller MobileNet classifier trained on either the PAS or IHC gold standard images, and show the confusion matrices in Fig. 5.5. As expected, the classification accuracy was much higher for IHC images (F1 = 0.86) compared to the PAS images (F1 = 0.63), as object classes can be determined by assessing stain positivity in IHC rather than morphological features in PAS. Classification errors from the model trained on IHC images were mostly evenly distributed across the three classes. We report qualitative results of the IHC model in Fig. 5.3, showing that our trained convolutional classifier is robust to stain variability between images. This variability would have made it difficult to infer stain positivity based on image intensity values alone.

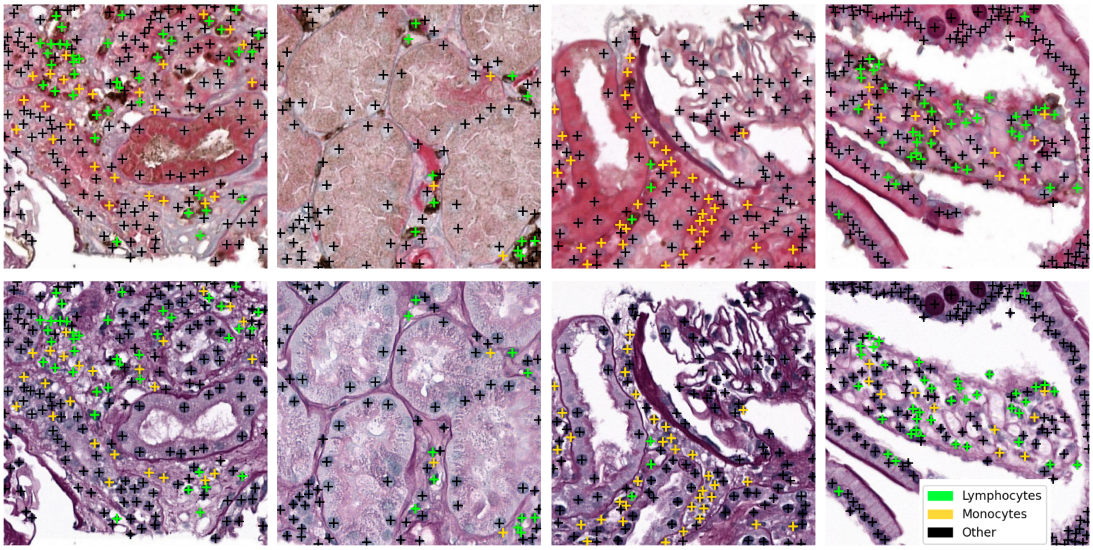

Figure 5.4: Qualitative results of the teacher model in the IHC stained images (top row), and label transfer to registered PAS images (bottom row). Note how the accurate registration between the two slides allows for cell to cell mapping between the two image modalities, which allowed for the creation of our silver standard dataset. It is also worth noting the difficulty of predicting cell classes from the PAS images alone.

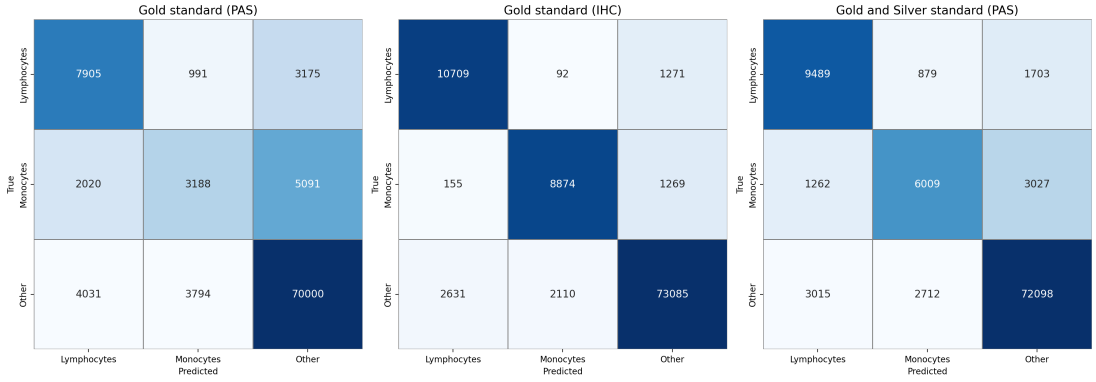

Figure 5.5: Confusion matrices showing the classification accuracy of the teacher model trained on the PAS and IHC gold standard datasets (left and centre) and the student model trained on gold and silver standard PAS images. From left to right, the corresponding F1 scores were 0.63, 0.86 and 0.75. All results are for the gold standard validation split.

	Task 1 (FROC)		**Task 2** (FROC)		
	Rank	Inflammatory Cells	Rank	Monocytes	Lymphocytes
InstanSeg-class (ours)	1	**0.42**	1	**0.44**	**0.20**
TIAKong	2	0.40	2	0.40	0.17
Aira Matrix	3	0.36	3	0.38	0.13

Table 5.1: Challenge results and ranking based on FROC on the public leaderboard. https://monkey.grand-challenge.org/evaluation/live-leaderboard/leaderboard/

Overall, the most difficult cell type to predict in the PAS images were the monocytes, which were also the minority class in the gold standard dataset. The relatively low F1 score of 0.63 is an indicator of the difficulty of the classification task. We demonstrate the classification difficulty in Fig. 5.2, by showing 16 examples of monocytes, lymphocytes and a hidden class. Stereotypically, monocytes should have a large ellipsoidal nucleus that can have lobules or dents, causing a bean-shaped appearance, while lymphocytes are much smaller cells with a darker and rounder nucleus. However, these three-dimensional structures can be lost during sample preparation due to tissue slicing and two-dimensional imaging. As a result, the majority of the lymphocytes and monocytes do not have canonical appearances in these images.

In Fig. 5.4, we show the classification results of our IHC model and the label transfer to the registered PAS-stained images. Note how the registration process enables one-to-one mapping of cells from one image modality to the other, justifying our approach of creating a silver-standard dataset. We show the classification results of our student model trained on both the gold and silver standard datasets in Fig. 5.5 (right panel). The classification accuracy of our teacher model was substantially higher (F1 = 0.75) than the model trained only on gold standard images (left panel), with the biggest improvement being for monocyte classification.

We obtain FROC scores by submitting a Docker contained pipeline on the Grand Challenge portal, which meant that the final test images were hidden to us. We report the FROC scores for overall inflammatory cell detection, as well as the individual scores for monocyte and lymphocyte detection in Table 5.1 and Table 5.2. Our method came first for both tasks on the public leaderboard which allowed for multiple submissions, and second in Task 1 and first in Task 2 for the final private leaderboard. Overall, our model performed especially well for the detection of monocytes compared to the other top performing methods in the competition.

	Task 1 (FROC)		**Task 2** (FROC)		
	Rank	Inflammatory Cells	Rank	Lymphocytes	Monocytes
InstanSeg-class (ours)	2	0.3875	1	0.4515	**0.2626**
TIAKong	1	**0.3930**	2	**0.4624**	0.2392
Aira Matrix	3	0.3517	3	0.4471	0.1906
Ourdiology	4	0.2861	4	0.3717	0.1573
ImmunoZip	5	0.2510	5	0.3064	0.0699

Table 5.2: Final challenge results and ranking based on FROC on the private leaderboard. https://monkey.grand-challenge.org/evaluation/final-test-phase/leaderboard/

5.5 Discussion

The MONKEY challenge received a total of 547 participants and 401 method submissions from 50 countries. Our solution consistently placed among the top two teams in the official challenge rankings, showing that our training method involving a large weakly annotated dataset of PAS images paired with a large EfficientNetV2 encoder allowed for state-of-the-art detection and classification of immune cells in PAS images.

Interestingly, our method was the only submission in the top performing teams that used a two-step process of cell segmentation followed by cell classification. Other teams instead chose pixel-level semantic segmentation (TIAKong) or recent object detection methods YOLO and DETR (Aira Matrix). Surprisingly, the two other challenge winners did not leverage the large unlabelled IHC images; instead these teams focused on mitigating class imbalances, augmentations and custom postprocessing steps. The TIAKong team also developed a surrogate training target that involved the semantic segmentation of cell classes and their boundaries to improve monocyte detection, which may have reduced over-fitting.

Our method leveraged the comparatively cheap cost of accurate WSI registration for substantially expanding the number of training samples. However, for tissue regions where cells are densely distributed, registration errors as small as 5 microns ($\approx$ 20 pixels) may cause the label of one cell to be incorrectly attributed to a neighbouring cell. As such, our method is expected to be highly sensitive to the accuracy of the upstream registration process. For translating this method to other datasets, we recommend an average registration error of less than a typical cell diameter. This requirement may prevent our method from being used on poorly fixed or fragile tissue sections, where accurate registration can be challenging using existing algorithms.

Our method was fundamentally limited by the ability of the InstanSeg model to detect all inflammatory cells in the test ROIs. Furthermore, because our instance segmentation method uses a seed map rather than an explicit foreground probability map, InstanSeg does not allow for the prediction of object existence probabilities. This has two main limitations, (1) cell class probabilities predicted by our classification models could not be combined with the probability of the cell existing, making our FROC metrics inexact, (2) InstanSeg could not be forced to include highly unconfident cells in its predictions, as lowering the seed detection threshold to arbitrarily low values substantially degrades overall segmentation accuracy. This capped the maximal recall that our method could reach. Unfortunately, the FROC score heavily penalises low recall at low detection thresholds, which may have disadvantaged our approach. Future improvements to InstanSeg could include the prediction of a separate object existence probability, which could be trained end-to-end with the classification model, and thereby allow for more accurate FROC scores and obtain arbitrarily high recall.

Our method was the only one among the five highest-ranked teams that did not use full image resolution during training or inference. Our choice of halving the resolution of the input images from 0.24 to 0.5 microns, allowed for the quadrupling of the number of training samples for the same I/O and storage costs. Early results suggested this led to higher accuracy than training at full resolution with a quarter of training examples. Furthermore, running inference at half resolution is likely to have substantially increased the inference speed of our method, although computational efficiency was not prioritised in this competition and timing information was not shared by the challenge organisers.

5.5.1 Usability and reproducibility

As for other work presented in this thesis, we recognise the importance of broader usability and reproducibility for both computer vision and biological research. Our code and trained models are open-source and available on GitHub[3], we also share a fully containerised version in Docker to ensure future reproducibility. We recognise that our trained models will be of greatest applicability to the community surrounding the challenge organisers and sponsors. However, because the MONKEY challenge involved six independent pathology institutes, it is likely that our findings and models can generalise to other research centres. Our method also provides a strong foundation for future adaptations to broader histopathological applications.

3. github.com/ThibautGoldsborough/instanseg-monkey-challenge

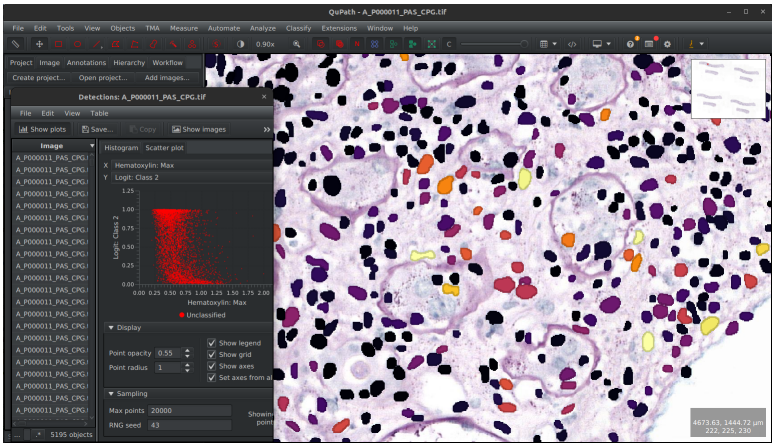

Figure 5.6: Screenshot showing the interactive InstanSeg extension in the QuPath software. Here, our InstanSeg model is paired with an EfficientNetV2 classifier to detect monocytes and lymphocytes. Monocyte probabilities are shown as an overlay, enabling the method to be integrated in full analyses pipelines with no coding experience required.

Our work aims to achieve more than a *proof-of-principle* that weakly supervised training enables high classification accuracy without requiring laborious and error prone expert annotations. We also gave attention to computational and implementation considerations. In an effort to improve our contribution to the bioimage community, we show that the InstanSeg model, object crop generation and object classification can all be serialised end-to-end in Torchscript and packaged in BioImageIO format (Ouyang et al., 2022) for inference in the software QuPath (Bankhead et al., 2017), as shown in Fig. 5.6.

Future research in this field, which could take place under the form of another challenge, should focus on the collection and curation of a broader range of weakly supervised tasks for training generalised cell classifiers in histopathology. Accurately registered WSIs, such as the ones provided in this challenge, can provide a huge number of labelled training pairs which could easily surpass the size of the largest supervised datasets of natural images. The difficulty will be to ensure diversity of tissue types, pathologies, staining techniques and image acquisition methods while ensuring that all the images have permissive licenses for re-use.

5.6 Conclusion

We have proposed a novel method for the detection and classification of inflammatory cells for the histopathological assessment of PAS-stained kidney biopsy images. Our approach involved a modular segment-then-classify pipeline that used an InstanSeg model for the unspecific segmentation of nuclei followed by a large CNN for the classification of small image patches centred on each nucleus. Our training paradigm leveraged the accurate image registration between IHC and PAS slides to transfer phenotypic information of nuclei from one modality to the other, rather than relying exclusively on the manually-annotated ground truth. Specifically, our method involved training a small teacher classifier on IHC data using manually annotated ground truth, which we then used to generate over two million weak annotations of registered PAS images. We used these weak annotations to train a larger student model using only PAS images. This paradigm yielded second place for overall inflammatory cell detection (0.3875 FROC) and first place for monocyte/lymphocyte classification (0.4515/0.2626 FROC) in the final MONKEY challenge rankings. Our approach demonstrated that weakly supervised label transfer can close some of the gap between non-specific PAS staining and immune-cell specific IHC staining references. Our method is fully open-source, containerized, and directly deployable in QuPath via BioImageIO compatible models, supporting reproducible research and the broader field of clinical pathology.

Chapter 6

Marker-agnostic cell phenotyping in multiplexed fluorescence images

6.1 Abstract

Multiplexed fluorescence imaging enables extensive, spatially resolved measurement of dozens of protein markers. These protein markers are indicative of cell type and function, which play an important role in both healthy and diseased tissues. To analyse such data, the boundaries between cells need to be determined and the specific combination of markers that are expressed in each cell needs to be assessed. In this chapter we (1) quantify how upstream cell segmentation affects downstream cell clustering and classification, and (2) introduce a novel, efficient method for predicting per-cell marker positivity that integrates local image features with global cell population context. We compare three segmentation methods (InstanSeg, Mesmer, CellposeSAM) on two public datasets (NaroNet, CPDMI) and show that InstanSeg consistently produces mean cell intensity profiles that allow for better clustering and improved classification. Building on this, we propose a MobileNet–ISAB architecture that combines a MobileNet encoder for local feature extraction with an Induced Set Attention Block (ISAB) transformer to incorporate global set-level information. We trained our model only on synthetic images and show real-world generalisation. Evaluated against Nimbus, a large U-Net based pixel classifier trained on 200 million annotations, our lightweight MobileNet–ISAB produces competitive results, at a fraction of the annotation cost. Our method is also computationally efficient, running up to three times faster than Nimbus.

6.2 Introduction

Multiplexed imaging methods now allow for the simultaneous, spatially resolved measurement of dozens or even hundreds of protein markers at subcellular resolution in tissue. These protein markers help us categorise cell types and understand cell functions, which is necessary to characterise the cellular organisation of tissue and investigate cell-to-cell interactions. The study of multiplexed images has allowed us to understand the cellular organisation of tissue (Hickey et al., 2023), the composition of tumours (Jackson et al., 2020) and tissue responses to infection (Delorey et al., 2021). However, the development of modern microscopy imaging platforms has been accompanied by a number of computational challenges for the reliable and scalable analysis of these images.

Like many bioimage analysis pipelines, multiplexed image analysis typically follows three steps: the detection of objects (usually cells), the extraction of meaningful features from these objects, and the classification or clustering of objects based on these extracted features.

The detection step is called segmentation and has received considerable attention in the field. Deep-learning based methods, including Mesmer (Greenwald et al., 2022), Cellpose (Stringer et al., 2021) and InstanSeg have demonstrated high cell segmentation accuracies in multiplexed fluorescence images, perhaps even matching human experts. In a previous chapter, we have shown that InstanSeg could be extended to use any number of imaging channels, to better capture the boundaries of cells based on the differential expression of biomarkers between adjacent cells. While improved cell segmentation should facilitate the downstream tasks of cell feature extraction and cell phenotyping, few studies have examined the impact of segmentation accuracy on downstream tasks.

In many studies, cell features are determined by extracting intensity measurements (e.g. mean, median, or maximum pixel values within a cell) or morphological measurements (e.g. circularity, area). Popular platforms such as QuPath (Bankhead et al., 2017) and CellProfiler (Stirling et al., 2021) facilitate the extraction of a number of these features for each cell and for each imaging channel. Many recent methods such as ASTIR (Geuenich et al., 2021), STELLAR (Brbić et al., 2022) and MAPS (Shaban

et al., 2024) still mainly use mean cell intensities across channels for representing cell features. The reason is partly due to field convention, as cell intensity expression matrices resemble single-cell flow cytometry data, a long time gold-standard for identifying cell types in biological samples.

However, the use of mean cell intensities in imaging data has its limitations. The mean cell intensity of a cell can be affected by neighbouring cells, due to spillover, blur, cellular protrusions, overlap between cells, staining artefacts and segmentation mistakes. In fact, studies have shown that over 20% of detected cells can have biologically implausible phenotypes due to these artefacts (Hunter et al., 2024). More powerful methods for extracting cell features involving deep neural networks (DNNs) have been proposed, including CellSighter (Amitay et al., 2023) and CelloType (Pang, Roy, Wu, & Tan, 2025). However, despite achieving high accuracies on selected test sets, the methods need to be retrained for images with different biomarker compositions. This severely hinders their use in real-world biological workflows.

There remains a number of difficulties impeding the development of a universal DNN for the automated extraction of cell features in multiplexed images. Firstly, most studies focus on downstream cell type classification, which inherently depends on biomarker composition, which in turn varies from image to image.

For example, a cell that is positive for the CD45 marker (denoted CD45+) would be classified as a leukocyte, but if it is also CD20+, it would be further classified as a B-cell, otherwise if it is CD3+ then it would be a T-cell, which can further be classified as a cytotoxic T-cell (CD8+), a helper T-cell (CD4+), or a regulatory T-Cell (CD4+ and FoxP3+), which we illustrate in Fig. 6.1. In general, specific combinations of markers indicate different cell types, but also cell states (e.g. proliferating cells should be Ki67+). Some combinations are generally not expected to be seen biologically (e.g. CD4+ and CD8+). Due to the close proximity of cells in tissue, small segmentation errors can lead to neighbouring cells being wrongly assigned as positive for a particular marker. Such errors can easily result in the wrong cell types or states being assigned, and skew downstream analyses.

Furthermore, a universal cell classifier in multiplexed images would require the names of the biomarkers but also biological context such as tissue type or pathology. One solution to this problem is to initially treat biomarkers independently, determine whether each cell is positive or negative for each marker, and then combine this information with marker metadata such as channel names (e.g. CD4, CD8 etc...) to obtain cell

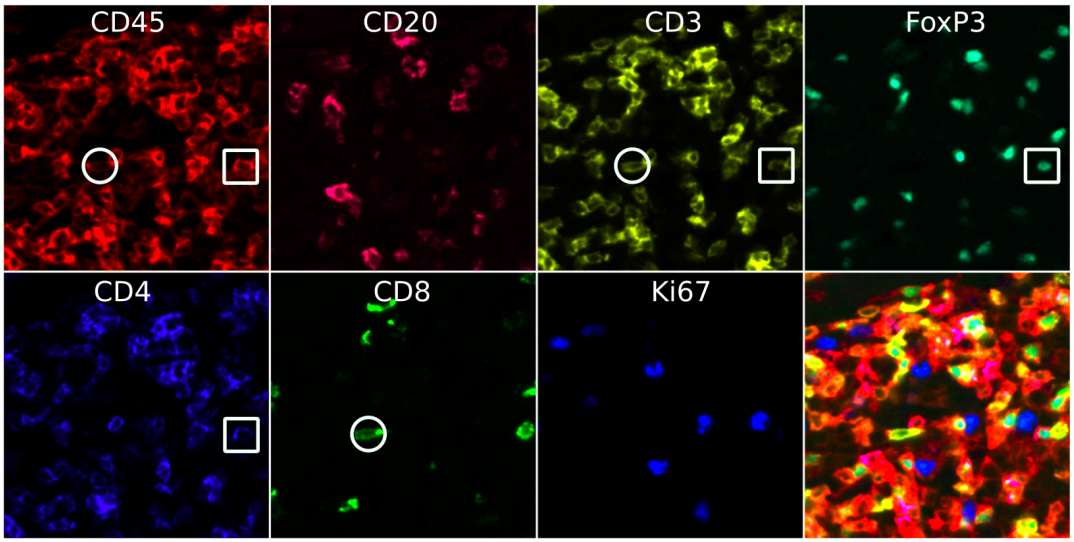

Figure 6.1: Example image showing how multiplexed images can reveal the cell type composition of tissue. The circled cell is CD45+, CD3+ and CD8+ (consistent with a cytotoxic T cell), while the squared cell is CD45+, CD3+, CD4+ and FOXP3+ (consistent with a regulatory T cell). Original fluorescence data from (J.-R. Lin et al., 2023).

types. This second step can be done by an informed biologist, and may be automated using large language models (LLMs) (X. J. Wang et al., 2024). However, it is the first step of predicting cell biomarker positivity that is both the most important and the most difficult.

One promising approach is Nimbus (Rumberger et al., 2024), which applies a U-Net to each channel independently to predict cell positivity according to a reference segmentation map. Nimbus was trained on the large Pan-M dataset, containing nearly 200 million cell annotations. However, the Pan-M annotations were generated semi-automatically, mostly using clustered or thresholded mean cell intensities as ground truth. This practice artificially inflates the size of the dataset without adding meaningful annotation information and may encourage shortcut-learning by models. For example, this may encourage models to learn dataset-specific annotation heuristics (such as intensity thresholds or clustering rules) rather than more relevant or robust imaging features. Furthermore, the use of a U-Net architecture bears two further limitations. Firstly, determining marker positivity requires comparison to the entire cell population. In whole slide images (WSIs), a negative cell population may be separated from a positive cell population by distances that far exceed the field of view (FOV) of a U-Net. In theory whole-slide normalisation provides some information of global cell

population context, however there is not yet a consensus in literature of how to normalise large WSIs consistently and efficiently. Global normalisation at full resolution is often prohibitive for large fluorescence images, while downsampling the image prior to normalising may be unreliable as values are likely to be influenced by both cell density and cell intensities. Tile-wise normalisation can introduce artefacts. Furthermore, bright extracellular off-target staining can cause the positive cell populations to have weak relative expression, which could be confused by a model with a small FoV as background or autofluorescence expression. Secondly, the Nimbus U-Net predicts marker positivity for each pixel, rather than each cell, which complicates downstream processing when pixel-wise predictions are not homogeneous within each cell.

Therefore, there remains a need for an improved and generalised computational method that can accurately phenotype segmented cells in multiplexed images. In a first part, this chapter assesses the effect of the cell segmentation method on downstream tasks such as clustering and classification using two publicly available datasets. In a second part, we introduce a novel method for the prediction of cell marker positivity. Our method combines the use of a convolutional network for extracting local image features and an Induced Set Attention Block (ISAB) transformer (Lee et al., 2019) for efficiently incorporating global cell population information. We train our method using only synthetic data and evaluate it on a small but diverse dataset of real multiplexed images.

6.3 Methods

6.3.1 Proposed method

Intuition

Mean cell intensity is a useful metric that is often used for predicting cell positivity, typically using a single cutoff threshold. However, automatically finding the right threshold value can be difficult, as it depends on illumination and contrast, as well as antibody affinity and marker location. These vary from marker to marker, image to image and experiment to experiment. Furthermore, even when an optimal threshold is determined, it rarely allows for the perfect separation of positive and negative cells. This is due to slight mistakes in segmentation, signal spillover from adjacent cells or weak membrane expression. More complex imaging features are required for this task.

Hence, the problem is in two parts: (1) the expression profiles of individual cells need to be assessed to quantify each cell's expression profile relative to its neighbours, and (2) the overall intensity distribution of positive and negative cells needs to be examined to determine which population of cells is positive (if any). The first of these requires local image features, while the second requires global image features.

As such, we also approach the task in two steps. In a first step, we use a convolutional classifier which takes as input a small image crop of a single channel and a binary marker of a cell of interest and outputs a vector embedding. In a second step, we use a transformer block which takes a bag of embeddings and outputs a logit predicting cell positivity for each embedding. This is similar to Multi Instance Learning (MIL) (Herrera et al., 2016), except instead of predicting a single logit for each bag, we predict a logit for each instance.

Some fluorescence images contain as many as a million cells, and the complexity of a standard transformer block $\mathcal{O}(n^2)$ would make it prohibitive. Instead, we use the Induced Set Attention Block (ISAB) from set transformers (Lee et al., 2019) defined as

$$\begin{aligned} \text{ISAB}(X) &= \text{MAB}(X, H, H) \in \mathbb{R}^{n \times d}, \\ \text{where} \quad H &= \text{MAB}(I, X, X) \in \mathbb{R}^{m \times d}, \end{aligned}$$

and $\text{MAB}(Q, K, V)$ is a Multihead Attention Block that computes attention between queries Q, keys K and values V. I is composed of m d-dimensional vectors, referred to as inducing points, which are trainable parameters learnt by the model that are expected to capture global properties of a distribution. For example, the m inducing points may be appropriately distributed points in embedding space, so that elements of a query may be summarised by their proximity to these inducing points. Of note, ISAB has a much lower complexity of $\mathcal{O}(nm)$ for m inducing points. This architecture also has the advantage of being equivariant to permutations, and accepts any number of items in each set.

Our method

Let $X_i \in \mathbb{R}^{2\times H\times W}$ denote the i-th single-cell image crop (one fluorescence channel and one binary cell mask), and let n be the number of cells in the bag. Our model components are:

$$\phi_\theta : \mathbb{R}^{2\times H\times W} \to \mathbb{R}^d \quad \text{(Image encoder)}$$

$$\mathrm{ISAB}_\psi : (\mathbb{R}^d)^n \to (\mathbb{R}^1)^n \quad \text{(Induced set attention block)}$$

Given a bag $(X_1, \ldots, X_n)$, the full model is

$$(\ell_1, \ldots, \ell_n) = \big(\mathrm{ISAB}_\psi \circ \phi_\theta^{(n)}\big)(X_1, \ldots, X_n),$$

where $\phi_\theta^{(n)}$ denotes applying ϕ_θ independently to each X_i.

The computational complexity of ISAB_ψ is $\mathcal{O}(nmd)$ for m inducing points and embedding size d. In practice, we use four attention heads, $m = 16$ inducing points, $d = 64$ embedding size, and train the model from scratch.

We instantiate ϕ_θ as a MobileNetV4 *small* (Qin et al., 2024) with pre-trained weights on Imagenet (Deng et al., 2009). We halve the stride of the first convolution from two to one to accommodate for small image sizes. Our combined model contained approximately 2.6 million parameters.

Training details

Throughout this chapter, we use image patches of size 64×64 at $0.5\,\mu\mathrm{m}$. We train the model end to end on the synthetic dataset, with batch size (number of sets) of 60 and set size of 100. We train over 250 epochs. We use the Adam optimiser (Kingma & Ba, 2017) with learning rate of $1e-3$ and weight decay of $1e-4$. We use focal loss (T. Y. Lin et al., 2017) instead of cross-entropy, which led to a minor performance increase (not shown). Focal loss puts less weight on easy to classify examples, which can be advantageous when the majority of examples are trivial to classify.

Ablations

To test the contribution of the two components of our method, we carry out an ablation study. First we replace the MobileNet with a conventional feature extractor that computes the mean, maximum, minimum and standard deviation of the pixel values within a cell mask to create a feature vector of length four for each cell. These features serve as input to our ISAB block. Separately, we test the contribution of the ISAB block by replacing it with a fully connected layer that processes each cell embedding in isolation.

6.3.2 Datasets

NaroNet (Jiménez-Sánchez et al., 2022)

We use the NaroNet High-grade Endometrial Carcinoma dataset (Jiménez-Sánchez et al., 2022), which we refer as NaroNet. The dataset comprises 360 images containing nearly two million cells. These images include tissue sections from twelve formalin-fixed, paraffin-embedded (FFPE) high-grade endometrial carcinomas, stained with a seven-colour multiplex panel encompassing DAPI, CD137, PD1, CD8, FOXP3, CD4 and CK. This dataset has the particularity that no current public cell segmentation method has been trained on this data. We use this dataset to compare publicly available segmentation methods on the downstream task of cell clustering.

CPDMI (Aleynick et al., 2023)

The Cross-Platform Dataset of Multiplex Fluorescent Cellular Object Image Annotations (Aleynick et al., 2023), which we refer to as CPDMI, is a publicly available dataset comprising multiplexed fluorescence images from a variety of human organs, including lung, breast, pancreas, colon, lymph node, ovary, skin, tongue, sacrum, hypopharynx, spleen, and tonsil. The images were captured using multiple imaging platforms, namely the Akoya Vectra, Zeiss Axioscan, and Akoya CODEX systems.

A small fraction of the cells in this dataset were manually segmented by the dataset authors, which was subsequently used to train the public InstanSeg *fluorescence nuclei and cells*[1] model. This likely incurs an advantage to InstanSeg when evaluated on this dataset. Other segmentation models we use to benchmark were not trained using these annotations, these methods instead used much larger training datasets with stricter conditions for re-use.

An experienced bioimage analyst (P. Bankhead) annotated by hand 1,000 cells across five markers and 20 fluorescence images totalling approximately 5,000 cell marker annotations. Each cell was annotated using a single point (typically towards the centre of the nucleus), and assigned a positive or negative classification for each marker of interest. This decision was made based only on the appearance of the cell within the image: biological knowledge of the marker was not used. Thus, a cell was annotated as positive for a marker if it *appeared* to be positive based on the information visible within that channel and the nucleus channel. Notably, this would mean that tumour infiltrating lymphocytes may be classified as cytokeratin positive because this would be the expected classification of a cell given only the cytokeratin channel and the nucleus. This decision was made to facilitate the development of a marker-agnostic image classification method that does not introduce a bias against uncommon biological phenotypes, while the interpretation of marker combinations may be treated as a subsequent step.

This is a somewhat subjective process: even by eye it is difficult to distinguish cell boundaries and there is no objective criterion by which to determine if the staining pattern and intensity is sufficient for a cell to be considered positive. The choice of cells to annotate was also subjective. However, this was an intentional decision to enrich the dataset for diverse morphologies and difficult but important cases - for example, neighbouring cells that are likely to be subject to false classifications due to signal spillover. A random selection of previously-segmented cells would be biased against cells that were poorly segmented or missed, while also failing to represent the most difficult aspects of the problem. Therefore despite the noisiness of the labels, we believe this dataset can serve well for validation by offering a unique challenge that aims to reflect real-world performance.

---

1. `https://github.com/instanseg/instanseg/releases/download/instanseg_models_v0.1.0/fluorescence_nuclei_and_cells.zip`

Synthetic Dataset

Generating ground truth annotations for multiplexed images is challenging and time consuming. Moreover, assessing cell positivity is often ambiguous or impossible for a large fraction of cells in an image. The few datasets that are publicly available have been obtained using automatic or semi-automatic methods such as thresholds or unsupervised clustering based on mean intensity features (Rumberger et al., 2024). This has several disadvantages: (1) it artificially inflates the size of datasets without adding meaningful annotation information, (2) it allows for shortcut learning by models, and (3) it undermines more advanced cell classification methods when evaluated on these datasets.

As such, we decide to generate synthetic data for training our models. We estimate that local image features at the single cell resolution are easier to synthesise than global image features such as tissue structures. More importantly, synthetic generation mitigates the ambiguity of assigning labels to each cell, which means that synthetic images can be annotated both densely and unambiguously.

Our synthetic generation pipeline takes as input a real instance segmentation of cells. For this we used the publicly available LuCa-7 image from Perkin Elmer [2] as well as a small crop of the CODEX WSI from CPDMI. Obtaining cell segmentation labels from real images is a hard requirement for this method, but these can be obtained cheaply using any of a range of publicly available cell segmentation models and public imaging data, circumventing the need for time consuming manual annotations. We stress that any input images could have been used for this step, as only the cell segmentation labels as opposed to the image data are used by synthetic generation pipeline. For each cell in the labelled map, we can generate a synthetic greyscale image representing a biomarker. We can control both the intensity and the subcellular location of the marker in the cell of interest, as well as the neighbouring cells.

Briefly, our pipeline is in two parts: first we use conventional image processing operations such as erosion, dilation and masking, to generate a binary mask of cellular expression in each cell. We mimic nuclear, cytoplasmic, cellular and membraneous expression, and we drop a random fraction of mask pixels to make the structures look

2. `https://downloads.openmicroscopy.org/images/Vectra-QPTIFF/perkinelmer/` available under CC-BY 4.0

realistic. In a second step, we generate a synthetic image from each binary mask. For this, we convolve the binary image with a Gaussian kernel to mimic the point spread function (PSF), and then use a mixed Poisson-Gaussian (MPG) process to mimic photon and detector noise,

$$y_{i,j} \sim \mathcal{P}(\lambda)x_{i,j} + \mathcal{N}(\mu, \sigma^2)$$

for each pixel location i, j, where y is output image, x is the image input, $\mathcal{P}$ is the Poisson distribution with gain λ, and $\mathcal{N}$ is the normal distribution with mean μ and standard deviation σ, a similar model was used for describing photon and detector noise in Foi, Trimeche, Katkovnik, and Egiazarian (2008). We show some examples of our synthetic images in Fig. 6.2.

We generate images in bags (or sets) to mimic crops taken from the same fluorescent channel. Within each bag, we keep marker expression to the same subcellular location. Interestingly, the mean cell intensity distribution of positive cells overlapped with the intensity distribution of negative cells, as shown in Fig. 6.2, which was an emergent property rather than one we hardcoded. This meant that the task of recovering cell labels could not be solved trivially, such as using a single threshold.

Notably, our pipeline is entirely written using batch parallelised Pytorch operations. On a laptop GPU, we could generate 25,000 sets of size 100 images in approximately 10 minutes. While very fast, this is too slow for on the fly generation, especially when a competing for GPU resources used for training. Hence, we precomputed these synthetic images and stored them in the *.h5* format for efficient downstream loading.

6.3.3 Metrics

Label-free segmentation metrics

Manually curated ground-truth annotations for cell segmentation are time-consuming, subjective, and often limited in size. Label-free metrics are a scalable alternative by quantifying how well extracted cell features form coherent, well-separated clusters in feature space. This approach, which has been used in other studies (Chen & Murphy, 2023; Zhu et al., 2024), is an indirect, but also less subjective and considerably cheaper alternative to using manually annotated references.

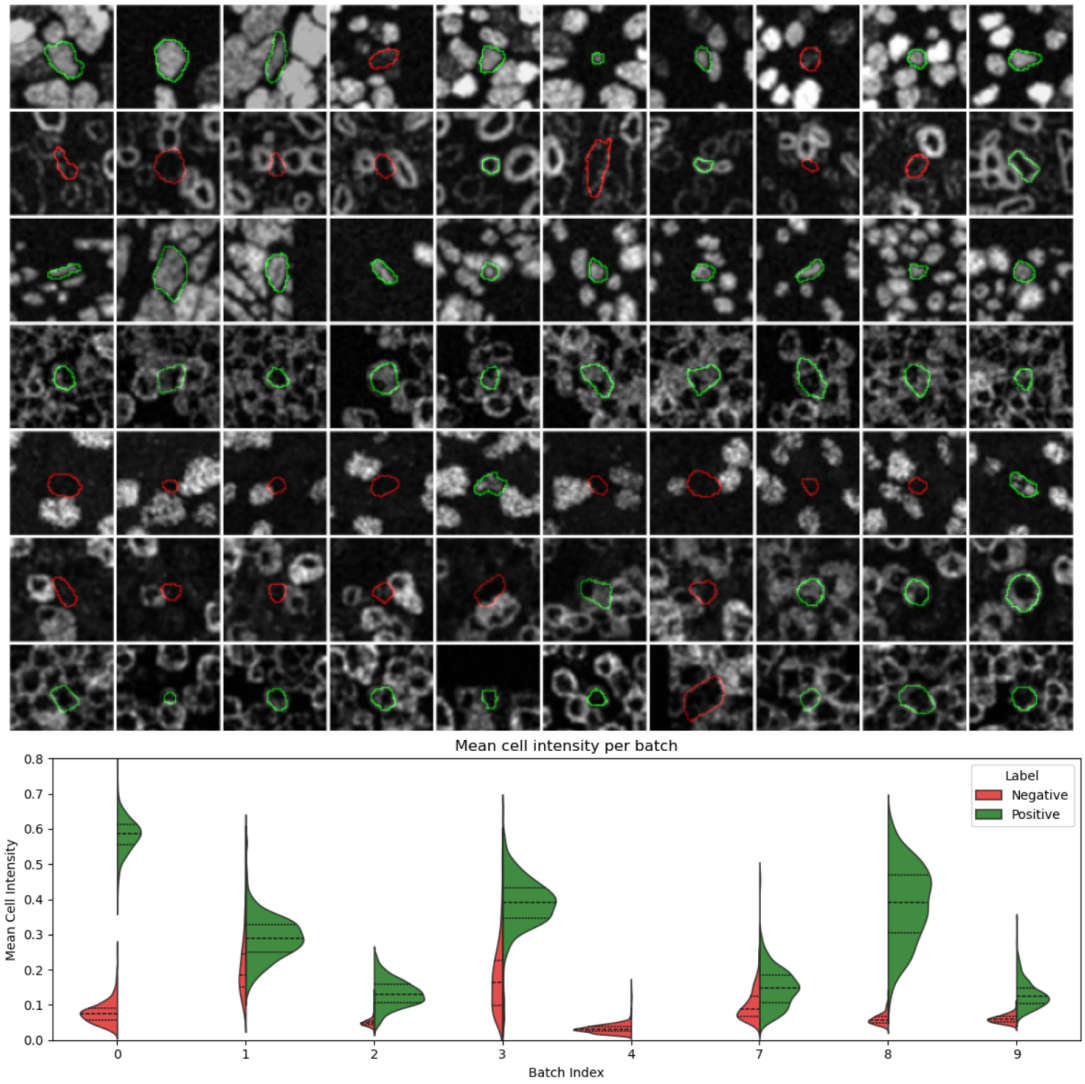

Figure 6.2: Example images and intensity distributions of from our synthetic dataset. Seven batches (rows) containing 10 images each. We show the outline of the cell of interest in green for positive cells and in red for negative cells. In the bottom plot, we show examples of the mean cell intensity distribution of the positive and negative cells (N = 1000 images for each batch). Note how the optimal threshold to separate the two cell populations differs from batch to batch. Note that the example images and mean cell intensity distributions in this figure are not paired.

For each of the N detected cells, we calculate the mean pixel intensity across C channels, which allows us to build a cell feature matrix of shape $N \times C$. We use the Scanpy library (F. A. Wolf, Angerer, & Theis, 2018) to z-normalise the features, build a neighbourhood graph, and cluster cell features using the Leiden algorithm (Traag, Waltman, & Van Eck, 2019) at three clustering resolutions: 0.05, 0.10 and 0.15. We project the cell features to two dimensions using the UMAP algorithm (McInnes, Healy, & Melville, 2018) for illustrative purposes. Fig. 6.3 shows the effect of increasing the clustering resolution on the number and size of observed cell clusters.

We use standard clustering metrics implemented in scikit-learn (Pedregosa et al., 2011). The Davies–Bouldin score quantifies the average similarity between each cluster and its most similar other cluster, with lower values indicating better separation. The Silhouette score measures how similar each sample is to its own cluster relative to other clusters, where higher values indicate more cohesive and well-separated clusters. The Calinski–Harabasz score evaluates the ratio of between-cluster to within-cluster dispersion, with higher values indicating better defined clusters. For each image, we compute these metrics at all three clustering resolutions and then average the results, producing a single value per metric per image.

On their own, these metrics have limited interpretive power, as they are only indirect measures of segmentation accuracy. However, we expect cells expression profiles of the same cell types to be similar. Mistakes in segmentation such as merging adjacent cells, or imprecise placement of cellular boundaries is expected to affect both the cohesion and separation of cell features. While absolute metric values may not be directly comparable across images, within a single image they should highlight subtle differences in segmentation accuracy.

6.3.4 Benchmarks

Segmentation benchmarks

For our segmentation benchmarks, we use some of the most widely used methods for the segmentation of cells in multiplexed images. We include Mesmer (Greenwald et al., 2022) which takes as input a two-channel image containing a nuclear marker and a cytoplasmic marker. We also include the recently released Cellpose SAM model (Pachitariu et al., 2025), which takes a three-channel input. On the NaroNet dataset,

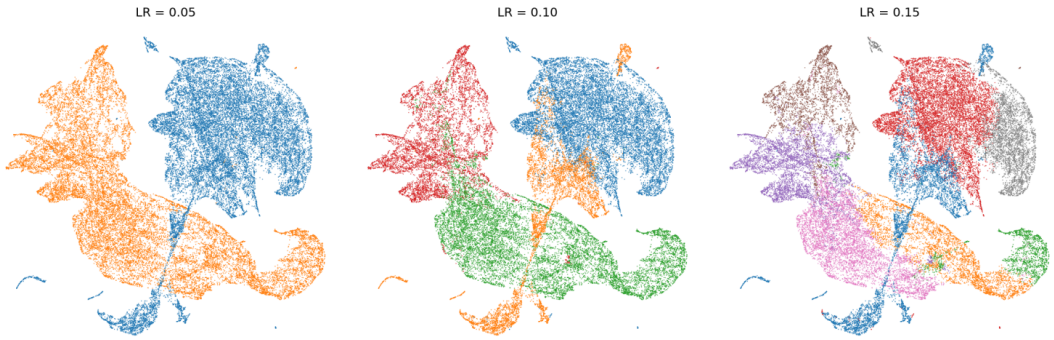

Figure 6.3: UMAPs with Leiden clustering of the mean cell features for increasing Leiden resolution (LR) using a single representative image of the NaroNet dataset segmented with InstanSeg. The unsupervised clustering algorithm produced 2,4 and 8 distinct cell clusters at 0.05, 0.10 and 0.15 Leiden resolutions, respectively.

we use the DAPI and CK markers as the nuclear and cytoplasmic markers, respectively. On the CPDMI dataset, which contains more diverse channel compositions, we use an aggregate of the non-nuclear markers to create the cytoplasmic channel (see Fig. 4.2 in Chapter 4).

Finally we include the publicly available InstanSeg fluorescence model [3] which accepts any number of channels as input. For all the models, we run segmentation at full image resolution (i.e. $0.5\,\mu\mathrm{m}$ / pixel) with default parameters.

Nimbus

We use the recently released Nimbus model (Rumberger et al., 2024) for the binary classification of cell-wise marker positivity. To the best of our knowledge, this is the only publicly available model for this task. Nimbus uses a large Residual U-Net architecture (approximately 36 million parameters) and was trained on the Pan-M dataset which contained 197 million annotations that were curated semi-automatically. Nimbus takes a two-channel image as input, one corresponding to a fluorescence channel and one corresponding to a binary map of cell segmentation, and outputs a binary semantic map of positive cells. To obtain a single prediction per cell, we average the pixel predictions under each cell mask.

3. https://github.com/instanseg/instanseg/releases/download/instanseg_models_v0.1.0/fluorescence_nuclei_and_cells.zip

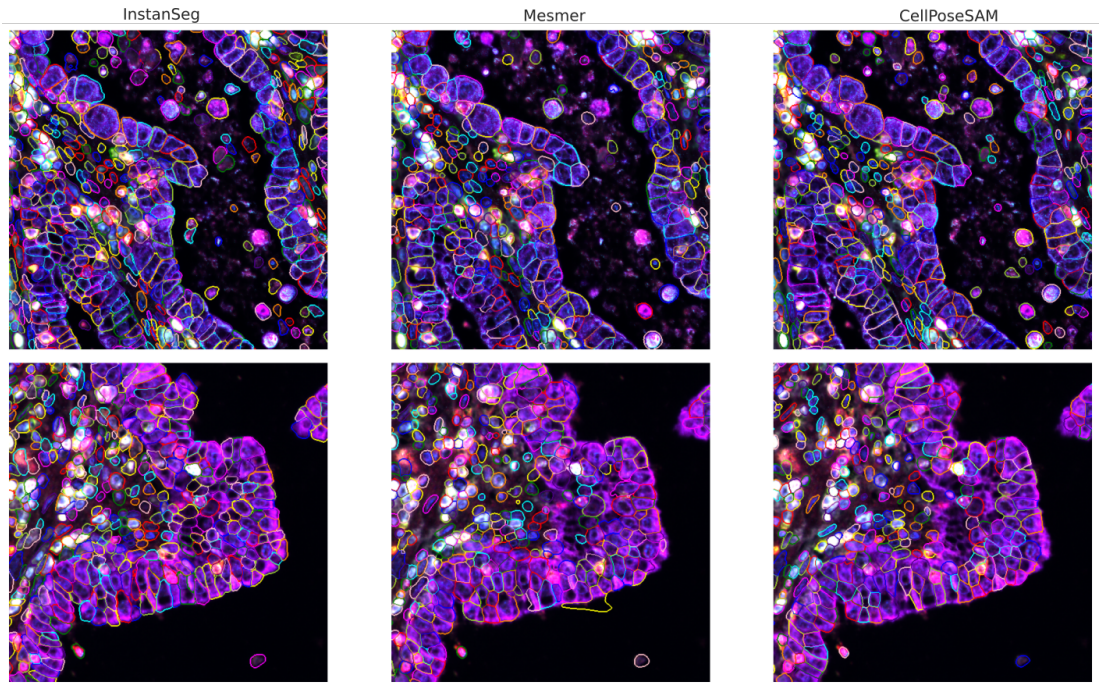

Figure 6.4: Qualitative examples of cell segmentation using InstanSeg, Mesmer and CellposeSAM on representative crops of the NaroNet dataset.

Mean cell intensity

We calculate the mean cell intensity (MCI) of each cell. To obtain an F1 score, we optimise the MCI threshold for each marker and each image so as to maximise the F1 score. This score is hence not a realistic measure of real world performance, but rather the best score that can theoretically be achieved using MCIs. This therefore represents an upper bound for the many methods that only use MCIs as input. It is likely also an upper bound for methods that predominantly rely on MCIs as input, such as ASTIR, MAPS and STELLAR.

6.4 Results

6.4.1 InstanSeg segmentation allows for better separation of cell clusters

We report qualitative segmentation results of InstanSeg, Mesmer and CellposeSAM on representative crops of the NaroNet dataset in Fig. 6.4, and illustrate in Fig. 6.5 a joint UMAP representation of the mean cell intensities. Overall, the three methods produced broadly similar segmentation maps and distributions of cell clusters.

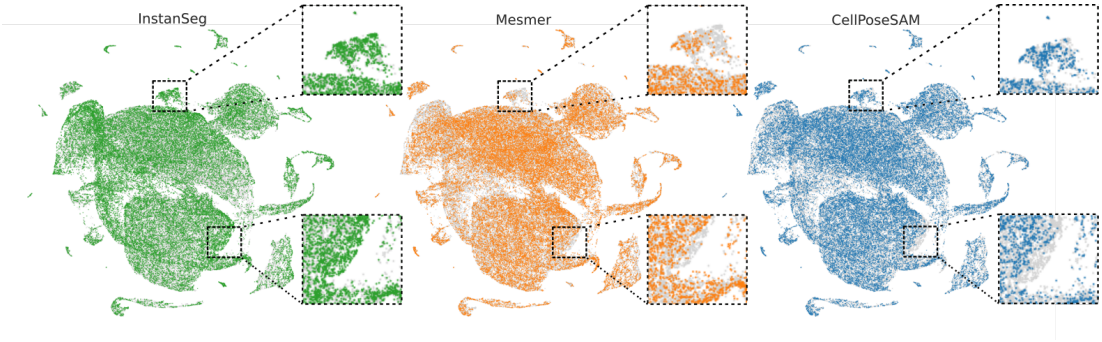

Figure 6.5: Joint UMAP representation of the mean cell intensities across seven image channels in the NaroNet dataset. We show cells from each segmentation method in colour over the combined distribution in gray. The arrows point to cell populations that were missed by some cell segmentation methods.

However, some cell populations, indicated by grey points in Fig. 6.5, were missed by either Mesmer or CellposeSAM but were mostly captured by InstanSeg. Conversely, InstanSeg did not miss populations detected by the other methods. While this observation alone is insufficient to conclude that the public InstanSeg fluorescence model produces superior segmentation, it suggests that the model may capture a broader diversity of cell types in the NaroNet dataset.

In Fig. 6.6, we report three commonly used metrics for evaluating cluster quality on the mean cell intensities obtained from the three methods. InstanSeg achieved the best performance on all metrics, with differences showing high statistical significance. These results suggest that InstanSeg may enable better separation of cell types, potentially indicating more precise placement of cellular boundaries.

6.4.2 InstanSeg paired with a MobileNet-ISAB classifier accurately predicts cell marker positivity

We report F1 scores for our ISAB cell classifier and Nimbus on the CPDMI dataset using InstanSeg, Mesmer, and CellposeSAM for upstream cell segmentation in Table 6.1. Overall, our ISAB classifier performed similarly to Nimbus. Our method produced better scores at the probability threshold of 0.5, while Nimbus produced better Area Under the Receiver Operating Characteristic Curve (ROC AUC). For two of the segmentation methods, InstanSeg and CellposeSAM, Nimbus and our

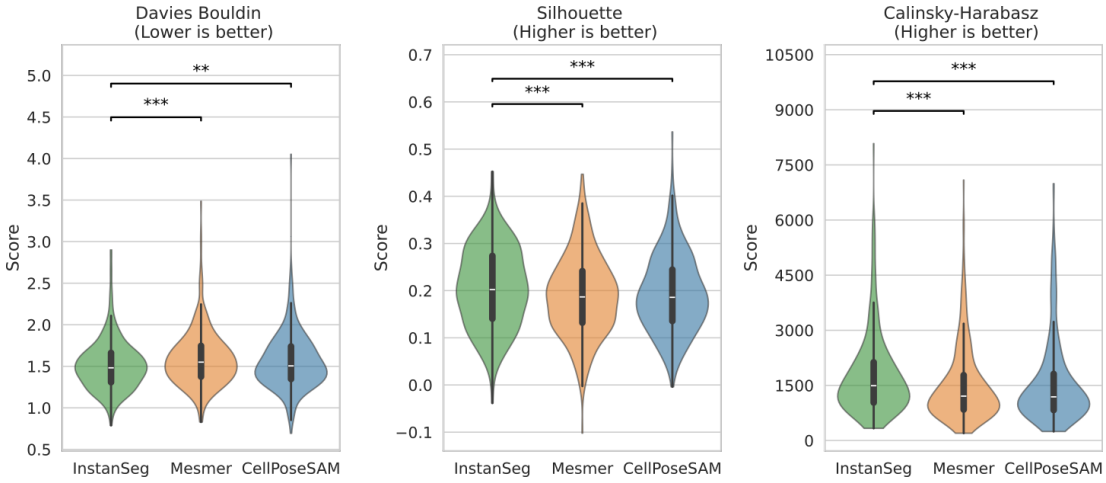

Figure 6.6: Clustering metrics of the mean cell intensity across seven image channels on the NaroNet dataset (N = 360). The metrics were calculated over Leiden clustering at 3 different Leiden Resolutions. Statistical test performed with the non-parametric Wilcoxon paired test, where we tested whether InstanSeg produced better clusters than the other methods.

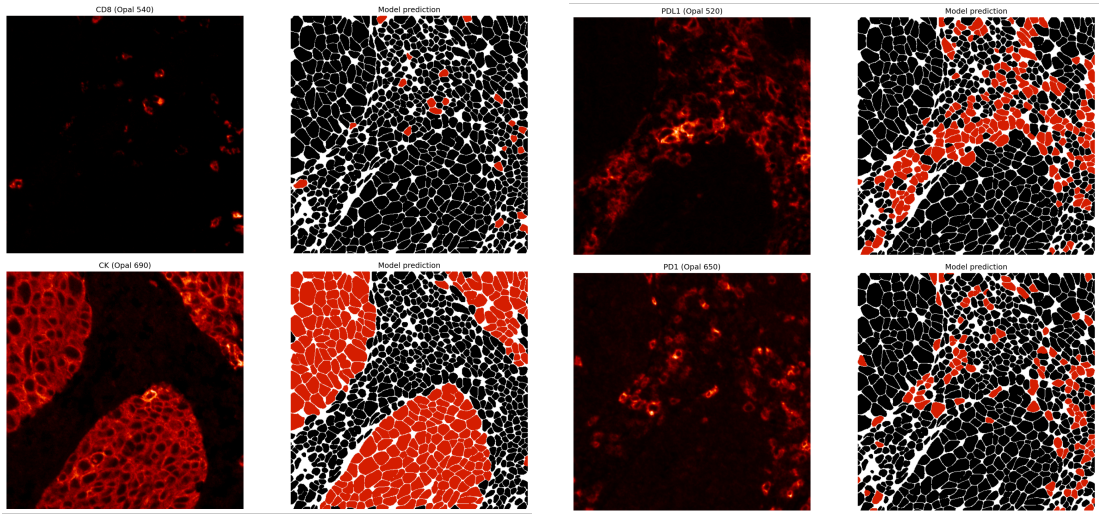

Figure 6.7: Qualitative prediction of marker positivity for four channels of the public LuCa7 image from Perkin Elmer. Cell segmentation was performed using all seven channels using the public InstanSeg fluorescence model, and cells were predicted as positive (red) or negative (black) for each channel using our ISAB classifier.

classifier outperformed the best attainable score when using mean cell intensities alone. All the classification methods produced higher scores when using InstanSeg for upstream cell segmentation, with the difference being especially notable for mean cell intensities.

Table 6.1: Performance metrics for cell classification methods and across segmentation methods. All methods were evaluated on our manually annotated CPDMI dataset, we omit cells that were not detected by the segmentation method in this analysis. Best F1 refers to the F1 at the optimal prediction threshold across all channels and images. For mean cell intensity, we optimise the cut-off threshold separately for each channel and image, and report the best achievable score.

	InstanSeg (N=5328)			Mesmer (N=5312)			Cellpose SAM (N=5088)		
	F1 @0.5	Best F1	ROC AUC	F1 @0.5	Best F1	ROC AUC	F1 @0.5	Best F1	ROC AUC
ISAB Classifier (Ours)	**0.886**	0.891	0.982	**0.855**	**0.857**	0.979	**0.874**	**0.881**	0.982
Nimbus	0.854	**0.898**	**0.985**	0.807	0.851	**0.979**	0.842	0.871	**0.983**
Mean cell intensity		0.885			0.868			0.874	

We report the results from our ablation studies in Table 6.2. Both the MobileNet and the ISAB blocks were necessary components of our architecture. Replacing ISAB with a fully connected layer that treats cell embeddings independently resulted in the biggest drop in accuracy, while replacing MobileNet with a simple conventional feature extractor led to a smaller but consistent decrease in performance.

We also present qualitative cell classification results of our ISAB classifier paired with upstream InstanSeg segmentation on a public multiplexed fluorescence image in Fig. 6.7. These results confirm that generalisation to real-world images can be achieved using only synthetic images for training. Notably, our method underperformed near image borders, an issue that could be mitigated in practice through an appropriate tiling strategy.

Table 6.2: Ablation study of our cell classification model on the manually annotated CPDMI dataset. 'No ISAB' replaces the Induced Set Attention Block with a fully connected layer that processes each cell embedding independently. 'No MobileNet' replaces the convolutional encoder with a feature extractor that computes the mean, minimum, maximum and standard deviation of the pixel values within each cell mask.

	InstanSeg (N=5328)			Mesmer (N=5312)			Cellpose SAM (N=5088)		
	F1 @0.5	Best F1	ROC AUC	F1 @0.5	Best F1	ROC AUC	F1 @0.5	Best F1	ROC AUC
ISAB Classifier (Ours)	**0.886**	**0.891**	**0.982**	**0.855**	**0.857**	**0.979**	**0.874**	**0.881**	**0.982**
No ISAB	0.436	0.445	0.733	0.386	0.422	0.687	0.424	0.446	0.731
No MobileNet	0.856	0.861	0.976	0.825	0.830	0.972	0.842	0.843	0.976

6.5 Discussion

The aim of this chapter was twofold. First, we investigated the effect of upstream cell segmentation method on downstream cell feature extraction. For this we investigated the overall distribution of mean cell intensities, a popular cell feature that is often used in the literature. Qualitatively, we found that InstanSeg may produce a more complete representation of the overall cell population in the large NaroNet dataset compared to some of the most widely used cell segmentation methods. In theory, spurious background detections by InstanSeg could also produce a similar result, however, we did not observe that InstanSeg detected entirely new cell clusters. Instead, InstanSeg provided a more complete representation of the clusters that were detected by either Mesmer and CellposeSAM.

Inspired by methods that could evaluate cell segmentation outputs without reference to ground truth annotations (Chen & Murphy, 2023), we compared cell segmentation methods on an image by image basis using three widely used clustering metrics. We found that InstanSeg segmentation produced mean cell intensities that could be better clustered than the other methods. While the absolute improvement in metrics was small, it was consistent across images, as evidenced by the high statistical significances. These metrics are imperfect, as they only indirectly relate to segmentation accuracy, however, unsupervised clustering is a common downstream task for cell type annotation and discovery. Thus, methods that produce better separation of cell types based solely on mean cell intensities remain valuable to the community.

Both Mesmer and CellposeSAM were trained on datasets with one to two orders of magnitude more annotations than the dataset used for our public InstanSeg model. Therefore, we attribute any observed improvements to InstanSeg's better handling of imaging channels (see Chapter 4), which in turn allows more precise placement of cellular boundaries, rather than to a larger or more diverse training set.

In the second part of this study, we compared methods for classifying cell positivity following upstream segmentation. We evaluated our ISAB-based cell classifier, trained solely on synthetic data, against Nimbus, a large U-Net pixel classifier trained with nearly 200 million annotations. Overall, Nimbus paired with InstanSeg achieved the highest F1 score (0.898) on our manual annotations. However, our smaller ISAB classifier surpassed Nimbus at a prediction probability cutoff of 0.5, suggesting better calibration.

When examining the effect of upstream cell segmentation on classification performance, we found that InstanSeg consistently outperformed Mesmer and Cellpose SAM. Furthermore, out of the 1000 dot annotations we made, InstanSeg only missed four cells, compared to seven for Mesmer and 50 missed cells for Cellpose SAM. We acknowledge that this comparison unfairly advantages InstanSeg as a small fraction of the CPDMI dataset was used to train the public InstanSeg model. Nevertheless, these results confirm that both Nimbus and our ISAB classifier generalise to other segmentation methods and benefit from improved segmentation. In fact, Nimbus performed better with InstanSeg than with Mesmer, despite Mesmer segmentations being prevalent in its Pan-M training data.

A key advantage of our training process is that we used only synthetic data and still achieved competitive performance on real-world images. Synthetic data offer several benefits: annotations are unambiguous and free from inter-annotator variability, and annotations can be extended to diverse cell morphologies, such as tissue-cultured cells. Moreover, synthetic datasets are easier to share. In contrast, at the time of writing, we could not download and process the full Pan-M dataset due to time and computational constraints.

Naturally, synthetic images have limitations. Designing a synthetic data generation pipeline requires domain expertise in microscopy and image processing, and there is a practical limit to the diversity of biomarkers that can be modelled using conventional image synthesis techniques. Our generation pipeline could be further improved; indeed, our classifier achieved an F1 score of 0.995 on our internal validation set, indicating that it may benefit from harder synthetic examples.

Future work should focus on combining real and synthetic image datasets, and could involve the large Pan-M dataset. Ideally, future annotations should cover a larger diversity of imaging modalities, perhaps even combining brightfield immunohistochemistry annotations with fluorescence ones. We are currently developing a new QuPath extension that will streamline the process of manual cell type annotation and proofreading, which should help us greatly expand the number, quality and diversity of our ground-truth annotations. Another valuable extension would be the prediction of continuous, or discrete, expression levels beyond binary positivity. This may be a task where synthetic data generation could be particularly useful, as this task is even harder for experts to annotate accurately.

From an architectural perspective, the ISAB transformer block enables a much larger field of view (FOV) than a U-Net. In fact, our method's FOV can span an entire whole-slide image, as processing one million cell embeddings of length 64 took under two seconds on a CPU and only 60 ms on a GPU. Furthermore, the MobileNet encoder was designed for runtime efficiency across a range of devices. Overall, despite scaling linearly with the number of cells, our method was roughly three times faster than the Nimbus U-Net.

It would be interesting to extend this architecture for the joint processing of image channels, possibly reusing elements of the architecture that was introduced in Chapter 4. Such an architecture could be used for the automatic channel-aware and population-aware clustering of cell types, and could incorporate a supervised *deep-clustering* training objective such as the one proposed by Brbić et al. (2022).

6.6 Conclusion

In this work, we investigated the impact of upstream cell segmentation on downstream clustering and classification in multiplexed fluorescence images and introduced a new method for predicting per-cell marker positivity. Across two public datasets, InstanSeg segmentation consistently produced mean cell intensities that allowed for more distinct clustering and improved classification performance compared with Mesmer and CellposeSAM. We further demonstrated that our novel, lightweight MobileNet–ISAB classifier, trained entirely on synthetic data, performed competitively with Nimbus, a large U-Net–based model trained on nearly 200 million annotations. Our approach generalised to real-world images without requiring fine-tuning, and our

synthetic training data benefits from both low annotation cost and easy portability. Together, these results demonstrate that InstanSeg paired with a downstream cell classifier can provide a practical and scalable solution for the analysis of the cellular composition of tissue, which we hope will support advances in biological discovery and digital pathology.

Chapter 7

Conclusion

7.1 Summary and Reflection

This thesis aimed to advance cell and tissue phenotyping by developing computational methods to automate and improve the analysis of microscopy images. Accurate phenotyping of cells is essential for understanding disease mechanisms, and supporting biological discovery. To address current limitations in accuracy, speed, and usability, we developed novel algorithms for the segmentation and classification of cells and nuclei in a variety of imaging modalities.

The first contribution of this thesis was the development of InstanSeg, an embedding-based instance segmentation algorithm that we initially optimized for the detection of nuclei in brightfield histology images. InstanSeg introduces a deep neural network to produce embeddings of image pixels that can then be efficiently and accurately clustered using a second neural network. InstanSeg consistently shows superior accuracy over other state-of-the-art segmentation methods when trained and evaluated on identical dataset splits using publicly available datasets.

A key advantage of InstanSeg is its lightweight postprocessing, which relies only on tensor operations. This dramatically accelerates postprocessing speeds compared to other segmentation methods. Faster processing speeds, combined with efficient tiling of microscopy images, allows for the high-throughput evaluation of large image collections, including whole-slide images which cannot fit in RAM, using widely available hardware. Another advantage of our tensor-based postprocessing is the ability to integrate InstanSeg in open-source software, without a requirement that it run within a research, Python-based environment. This allows our method to be used by biologists with no coding experience in user-friendly software, and readily integrated in full analysis pipelines.

The second major contribution of this thesis was the adaptation of InstanSeg for the joint segmentation of nuclei and cells in multiplexed fluorescence microscopy images. For this we introduced ChannelNet, a novel UNet-based architecture inspired by neural networks that could operate on sets. ChannelNet can produce three-channel representations of multiplexed images irrespective of the number and ordering of imaged biomarkers. When trained end-to-end with InstanSeg, we showed that our method set a new baseline for the segmentation of cells in multiplexed images. Notably, we found that segmentation accuracy improved as the number of imaging channels increased, demonstrating that ChannelNet could capture the boundaries of cells based on variations in cellular expression across entire biomarker panels. Improved segmentation accuracies in multiplexed imaging will allow for better downstream cell and tissue phenotyping.

An interesting alternative use case of our ChannelNet implementation was proposed in RNA2Seg (Defard et al., 2025) for the segmentation of cells using RNA-transcriptomics data, in which RNA molecules are spatially resolved as individual points. The authors used ChannelNet to predict cellular boundaries based on the unique transcriptomic profile of each cell. It was encouraging to see that our code base could be repurposed for alternative but related applications.

One limitation of the work presented in this thesis is that we have not extended our methods to 3D and 4D (3D + time) microscopy imaging. While extending our U-Net based encoders and embedding-based pixel clustering to further dimensions is conceptually trivial, ensuring that 3D models have similar generalisation performance of 2D models using very limited public 3D annotations is much more difficult. One avenue that was investigated was the use of ACS convolutions (Yang et al., 2021) for the adaptation of trained 2D models to 3D, which yielded early promising results and grounds for further investigation. However, we ultimately decided to focus on 2D applications (including highly-multiplexed images) as we judged that this represented an overwhelmingly more urgent need for the research community at this time. This reflects the current state of pathology imaging, where fully 3D data remain rare, costly, and far from standardised. For example, some research applications have explored the use of light-sheet microscopy for non-destructive tissue imaging (Bishop et al., 2024), whereas elsewhere serially-section images are acquired with H&E staining (Kiemen et al., 2022). Should any of these approaches become widespread, InstanSeg's performance and computational efficiency may make it an ideal starting point for the development of novel 3D cell segmentation methods in the future.

As of writing, our pre-trained models have been downloaded over 16,000 times, and our Python package has been installed 14,000 times. Our GitHub repository has received over 175 stars, suggesting that our code base has been of use to other researchers in the field. Excitingly, our method has been independently integrated in other open-source projects such as LazySlide (Zheng, Abila, Chrenková, Winkler, & Rendeiro, 2025), CellACDC (Padovani, Mairhörmann, Falter-Braun, Lengefeld, & Schmoller, 2022), Cuisto (Le Goc, d'Humières, Lesage, & Bouvier, 2025), MuSpAn (Bull, Moore, Mulholland, Leedham, & Byrne, 2024), Kintsugi (Smith et al., 2025) and DaneelPath (Vieco-Martí, López-Carrasco, Navarro, Granados Aparici, & Noguera, 2025). We hope that this will further decrease the barrier to using our method and increase the range of applications.

Most importantly, InstanSeg and its QuPath extension have demonstrated rapid adoption among biology researchers. It has been exceptionally rewarding to see that our models have already been used for cancer research, including melanoma (Guedes et al., 2025), B-cell lymphoma (Wikström & Hammer, 2025), (Hashemi, Prakash, Amini, Hollander, & Enblad, 2025), (Mears et al., 2025), triple negative breast cancer (Grevitt et al., 2025) brain neuroendocrine cancer (McNicoll et al., 2025), renal carcinomas (Toma et al., 2025) and glioblastoma (Kint et al., 2025). InstanSeg has contributed to mapping out the cellular and molecular landscape of muscle-invasive bladder cancer (Wahlin, 2025) and has helped reveal the distinct cellular interactions and immune micro-environments in diffuse large B-cell lymphomas (Qian & Zhang, 2025).

In another chapter, we showed that InstanSeg could be paired with downstream deep-learning classifiers for the accurate phenotyping of cells in brightfield images. We demonstrated this by developing a method that could accurately detect immune cell types in renal biopsies. For this, we made extensive use of registered immunohistochemistry images to transfer phenotypic information of immune cell types from one image modality to the other, without requiring extensive manual annotations. Our solution demonstrated state-of-the-art performance by coming joint first place in the public MONKEY leaderboard hosted on Grand Challenge. The high performance demonstrated that cell segmentation could be effectively decoupled from downstream cell classification. This modularity simplifies downstream applications and integration into full analysis pipelines.

In a last chapter, we focused on the automated phenotyping of cells in multiplexed fluorescence images. In a first part, we investigated the impact of cell segmentation on downstream cell feature representations, clustering and classification. We then introduced a transformer-based set classifier that could predict cell marker positivity based on both local and global image context. Excitingly, we could train our model using synthetic images only and obtain real-world generalisation performance.

We anticipate that improvements in registration between different imaging modalities, including both immunohistochemistry and multiplexed fluorescence images, will soon provide substantial training data for deep-learning classifiers for the phenotyping of cells in microscopy images, with the potential of surpassing the accuracy of domain experts.

In conclusion, this thesis has contributed novel tools for the automated analysis and quantification of microscopy images, focusing on the segmentation and phenotyping of nuclei and cells in both brightfield and fluorescence images. We ensured our methods could be used by biologists without coding experience, which has already enabled fast adoption for a wide range of research applications. We hope that the tools presented in this thesis will continue to support advances in biological discovery and digital pathology.

Bibliography

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... others (2016). Tensorflow: a system for large-scale machine learning. In *12th usenix symposium on operating systems design and implementation (osdi 16)* (pp. 265–283).

Acs, B., Ahmed, F. S., Gupta, S., Wong, P. F., Gartrell, R. D., Sarin Pradhan, J., ... Rimm, D. L. (2019). An open source automated tumor infiltrating lymphocyte algorithm for prognosis in melanoma. *Nature communications*, *10*(1), 5440.

Aleynick, N., Li, Y., Xie, Y., Zhang, M., Posner, A., Roshal, L., ... Hollmann, T. J. (2023). Cross-platform dataset of multiplex fluorescent cellular object image annotations. *Scientific Data*, *10*, 193. doi: 10.1038/s41597-023-02108-z

Amitay, Y., Bussi, Y., Feinstein, B., Bagon, S., Milo, I., & Keren, L. (2023, July). CellSighter: a neural network to classify cells in highly multiplexed images. *Nature Communications*, *14*(1), 4302. doi: 10.1038/s41467-023-40066-7

Amos, B. (2000, August). Lessons from the history of light microscopy. *Nature Cell Biology*, *2*(8), E151–E152. doi: 10.1038/35019639

Angelo, M., Bendall, S. C., Finck, R., Hale, M. B., Hitzman, C., Borowsky, A. D., ... Nolan, G. P. (2014, April). Multiplexed ion beam imaging of human breast tumors. *Nature Medicine*, *20*(4), 436–442. doi: 10.1038/nm.3488

Aubreville, M., Stathonikos, N., Bertram, C. A., Klopfleisch, R., Ter Hoeve, N., Ciompi, F., ... others (2023). Mitosis domain generalization in histopathology images—the midog challenge. *Medical Image Analysis*, *84*, 102699.

Avrameas, S., & Uriel, J. (1966). [Method of antigen and antibody labelling with enzymes and its immunodiffusion application]. *Comptes Rendus Hebdomadaires Des Seances De l'Academie Des Sciences. Serie D: Sciences Naturelles*, *262*(24), 2543–2545.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bai, J., Lu, F., Zhang, K., et al. (2019). *Onnx: Open neural network exchange.* `https://github.com/onnx/onnx`. GitHub.

Bancroft, J. D. (2019). 3 - light microscopy. In S. K. Suvarna, C. Layton, & J. D. Bancroft (Eds.), *Bancroft's theory and practice of histological techniques (eighth edition)* (Eighth Edition ed., p. 25-39). Elsevier. doi: https://doi.org/10.1016/B978-0-7020-6864-5.00003-7

Bankhead, P. (2022). *Introduction to bioimage analysis*. Online: bioimagebook.github.io. Retrieved from `https://bioimagebook.github.io/` (Interactive Jupyter-Book, CC-BY 4.0 license)

Bankhead, P., Loughrey, M. B., Fernández, J. A., Dombrowski, Y., McArt, D. G., Dunne, P. D., ... Hamilton, P. W. (2017, December). QuPath: Open source software for digital pathology image analysis. *Scientific Reports*, *7*(1), 16878. doi: 10.1038/s41598-017-17204-5

Baydin, A. G., Pearlmutter, B. A., Radul, A. A., & Siskind, J. M. (2018, February). *Automatic differentiation in machine learning: a survey.* arXiv. (arXiv:1502.05767 [cs]) doi: 10.48550/arXiv.1502.05767

Berg, S., Kutra, D., Kroeger, T., Straehle, C. N., Kausler, B. X., Haubold, C., ... Kreshuk, A. (2019, December). ilastik: interactive machine learning for (bio)image analysis. *Nature Methods*, *16*(12), 1226–1232. doi: 10.1038/s41592-019-0582-9

Berman, M., Triki, A. R., & Blaschko, M. B. (2018, June). The Lovasz-Softmax Loss: A Tractable Surrogate for the Optimization of the Intersection-Over-Union Measure in Neural Networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4413–4421). Salt Lake City, UT: IEEE. doi: 10.1109/CVPR.2018.00464

Bishop, K. W., Erion Barner, L. A., Han, Q., Baraznenok, E., Lan, L., Poudel, C., ... others (2024). An end-to-end workflow for nondestructive 3d pathology. *Nature Protocols*, *19*(4), 1122–1148.

Brbić, M., Cao, K., Hickey, J. W., Tan, Y., Snyder, M. P., Nolan, G. P., & Leskovec, J. (2022). Annotation of spatially resolved single-cell data with stellar. *Nature Methods*, *19*(11), 1411–1418.

Budelmann, D., Weiss, N., Heldmann, S., & Lotz, J. (2022). Histokatfusion. *Image registration for the ACROBAT challenge*.

Bull, J. A., Moore, J. W., Mulholland, E. J., Leedham, S. J., & Byrne, H. M. (2024). Muspan: A toolbox for multiscale spatial analysis. *bioRxiv*, 2024–12.

Caicedo, J. C., Goodman, A., Karhohs, K. W., Cimini, B. A., Ackerman, J., Haghighi, M., ... Carpenter, A. E. (2019, December). Nucleus segmentation across imaging experiments: the 2018 Data Science Bowl. *Nature Methods*, *16*(12), 1247–1253. doi: 10.1038/s41592-019-0612-7

Cardoso, M. J., Li, W., Brown, R., Moinuddin, S., Murrey, T., Zhao, S., ... others (2023). Monai: An open-source framework for deep learning in healthcare. *Computing in Science & Engineering*, *25*(2), 82–95.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In *European conference on computer vision* (pp. 213–229).

Carpenter, A. E., Kamentsky, L., & Eliceiri, K. W. (2012, July). A call for bioimaging software usability. *Nature Methods*, *9*(7), 666–670. doi: 10.1038/nmeth.2073

Chandrasekaran, S. N., Cimini, B. A., Goodale, A., Miller, L., Kost-Alimova, M., Jamali, N., ... others (2024). Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations. *Nature Methods*, *21*(6), 1114–1121.

Chen, H., & Murphy, R. F. (2023). Evaluation of cell segmentation methods without reference segmentations. *Molecular biology of the cell*, *34*(6), ar50.

Chen, H., Qi, X., Yu, L., & Heng, P.-A. (2016). Dcan: deep contour-aware networks for accurate gland segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2487–2496).

Cho, N. H., Cheveralls, K. C., Brunner, A.-D., Kim, K., Michaelis, A. C., Raghavan, P., ... others (2022). Opencell: Endogenous tagging for the cartography of human cellular organization. *Science*, *375*(6585), eabi6983.

Cifci, D., Foersch, S., & Kather, J. N. (2022). Artificial intelligence to identify genetic alterations in conventional histopathology. *The Journal of pathology*, *257*(4), 430–444.

Cutler, K. J., Stringer, C., Lo, T. W., Rappez, L., Stroustrup, N., Brook Peterson, S., ... Mougous, J. D. (2022). Omnipose: a high-precision morphology-independent solution for bacterial cell segmentation. *Nature methods*, *19*(11), 1438–1448.

Danial, J. S. (2025). Super-resolution microscopy for structural biology. *Nature Methods*, 1–17.

Dayao, M. T., Brusko, M., Wasserfall, C., & Bar-Joseph, Z. (2022, April). Membrane marker selection for segmenting single cell spatial proteomics data. *Nature Communications*, *13*, 1999. doi: 10.1038/s41467-022-29667-w

De Brabandere, B., Neven, D., & Van Gool, L. (2017). Semantic instance segmentation with a discriminative loss function. *arXiv preprint arXiv:1708.02551*.

Defard, T., Blondel, A., Bellow, S., Coleon, A., de Melo, G. D., Walter, T., & Mueller, F. (2025). Rna2seg: a generalist model for cell segmentation in image-based spatial transcriptomics. *bioRxiv*, 2025–03.

Delorey, T. M., Ziegler, C. G. K., Heimberg, G., Normand, R., Yang, Y., Segerstolpe, Å., ... other consortium authors (2021). Covid-19 tissue atlases reveal sars-cov-2 pathology and cellular targets. *Nature*, *595*(7865), 107–113. doi: 10 .1038/s41586-021-03570-8

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 248–255). doi: 10.1109/CVPR.2009 .5206848

Deotale, G., Ambast, A., Ramchandani, L., Das, D. K., & Thomas, T. (2025). Ensemble object detection methodology for automated detection of inflammatory cells in kidney biopsies. In *Medical imaging with deep learning - short papers.*

DeVito, Z., Ansel, J., Constable, W., Suo, M., Zhang, A., & Hazelwood, K. (2021). Using python for model inference in deep learning. *arXiv preprint arXiv:2104.00254*.

Dominguez Mantes, A., Herrera, A., Khven, I., Schlaeppi, A., Kyriacou, E., Tsissios, G., ... others (2025). Spotiflow: accurate and efficient spot detection for fluorescence microscopy with deep stereographic flow regression. *Nature Methods*, 1–10.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... Houlsby, N. (2021, June). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.* arXiv. (arXiv:2010.11929 [cs]) doi: 10.48550/arXiv.2010.11929

Edlund, C., Jackson, T. R., Khalid, N., Bevan, N., Dale, T., Dengel, A., ... Sjögren, R. (2021). Livecell—a large-scale dataset for label-free live cell segmentation. *Nature methods*, *18*(9), 1038–1045.

Fathi, A., Wojna, Z., Rathod, V., Wang, P., Song, H. O., Guadarrama, S., & Murphy, K. P. (2017). Semantic instance segmentation via deep metric learning. *arXiv preprint arXiv:1703.10277*.

Foi, A., Trimeche, M., Katkovnik, V., & Egiazarian, K. (2008). Practical Poissonian-Gaussian noise modeling and fitting for single-image raw-data. *IEEE Transactions on Image Processing*, *17*(10), 1737–1754. doi: 10.1109/TIP.2008.2001399

Franco-Barranco, D., Andrés-San Román, J. A., Hidalgo-Cenalmor, I., Backová, L., González-Marfil, A., Caporal, C., ... Arganda-Carreras, I. (2024). Biapy: Accessible deep learning on bioimages. *bioRxiv*. doi: 10.1101/2024.02.03.576026

Furness, P. N., & Taub, N. (2001, November). International variation in the interpretation of renal transplant biopsies: Report of the CERTPAP Project1. *Kidney International*, *60*(5), 1998–2012. doi: 10.1046/j.1523-1755.2001.00030.x

Geuenich, M. J., Hou, J., Lee, S., Ayub, S., Jackson, H. W., & Campbell, K. R. (2021). Automated assignment of cell identity from single-cell multiplexed imaging and proteomic data. *Cell Systems*, *12*(12), 1173–1186.e6. doi: 10.1016/j.cels.2021.08.012

Girshick, R. (2015, September). *Fast R-CNN.* arXiv. (arXiv:1504.08083 [cs]) doi: 10.48550/arXiv.1504.08083

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 580–587).

Glorot, X., Bordes, A., & Bengio, Y. (2011). Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics* (pp. 315–323).

Goldsborough, T., Philps, B., O'Callaghan, A., Inglis, F., Leplat, L., Filby, A., ... Bankhead, P. (2024, August). *InstanSeg: an embedding-based instance segmentation algorithm optimized for accurate, efficient and portable cell segmentation.* arXiv. (arXiv:2408.15954 [cs]) doi: 10.48550/arXiv.2408.15954

Goltsev, Y., Samusik, N., Kennedy-Darling, J., Bhate, S., Hale, M., Vazquez, G., ... Nolan, G. P. (2018, August). Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging. *Cell*, *174*(4), 968–981.e15. (Publisher: Elsevier) doi: 10.1016/j.cell.2018.07.010

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning* (Vol. 1) (No. 2). MIT press Cambridge.

Graham, S., Vu, Q. D., Raza, S. E. A., Azam, A., Tsang, Y. W., Kwak, J. T., & Rajpoot, N. (2019, December). Hover-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical Image Analysis*, *58*, 101563. doi: 10.1016/j.media.2019.101563

Greenwald, N. F., Miller, G., Moen, E., Kong, A., Kagel, A., Dougherty, T., ... Van Valen, D. (2022, April). Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. *Nature Biotechnology*, *40*(4), 555–565. doi: 10.1038/s41587-021-01094-0

Grevitt, P., Maiques, O., Shah, K. M., Morales, V., Gadaleta, E., Dodel, M., ... others (2025). Coenzyme a homeostasis regulates hypoxic signalling via ubfd1 in triple-negative breast cancer. *bioRxiv*, 2025–06.

Guedes, J., Szadai, L., Woldmar, N., Jánosi, Á. J., Koroncziová, K., Lengyel, B. M., ... others (2025). The melanoma mega-study: Integrating proteogenomics, digital pathology, and ai-analytics for precision oncology. *Journal of proteomics*, 105482.

Guerrero-Pena, F. A., Fernandez, P. D. M., Ren, T. I., Yui, M., Rothenberg, E., & Cunha, A. (2018). Multiclass weighted loss for instance segmentation of cluttered cells. In *2018 25th IEEE international conference on image processing (icip)* (pp. 2451–2455).

Haas, M., Loupy, A., Lefaucheur, C., Roufosse, C., Glotz, D., Seron, D. a., ... others (2018). *The banff 2017 kidney meeting report: Revised diagnostic criteria for chronic active t cell–mediated rejection, antibody-mediated rejection, and prospects for integrative endpoints for next-generation clinical trials.* Wiley Online Library.

Harris, C. R., Millman, K. J., Van Der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... others (2020). Array programming with numpy. *Nature*, *585*(7825), 357–362.

Hashemi, J., Prakash, S., Amini, R., Hollander, P., & Enblad, G. (2025). Immunoprofiling of the tumor microenvironment in dlbcl using multiplexed immunofluorescence and qupath analysis. *Hematological Oncology*, *43*, e494_70096.

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2018, January). *Mask R-CNN.* arXiv. (arXiv:1703.06870 [cs])

Hekler, A., Utikal, J. S., Enk, A. H., Solass, W., Schmitt, M., Klode, J., ... others (2019). Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images. *European Journal of Cancer*, *118*, 91–96.

Henriques, R., Griffiths, C., Hesper Rego, E., & Mhlanga, M. M. (2011). Palm and storm: unlocking live-cell super-resolution. *Biopolymers*, *95*(5), 322–331.

Herrera, F., Ventura, S., Bello, R., Cornelis, C., Zafra, A., Sánchez-Tarragó, D., & Vluymans, S. (2016). Multiple instance learning. In *Multiple instance learning: foundations and algorithms* (pp. 17–33). Springer.

Hickey, J. W., Becker, W. R., Nevins, S. A., Horning, A. M., Perez, A. E., Zhu, C., ... other consortium authors (2023). Organization of the human intestine at single-cell resolution. *Nature*, *619*(7970), 572–584. doi: 10.1038/s41586-023-05915-x

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Hu, M., Maillard, M., Zhang, Y., Ciceri, T., La Barbera, G., Bloch, I., & Gori, P. (2020). Knowledge distillation from multi-modal to mono-modal segmentation networks. In *International conference on medical image computing and computer-assisted intervention* (pp. 772–781).

Hunter, B., Nicorescu, I., Foster, E., McDonald, D., Hulme, G., Fuller, A., ... others (2024). Optimal: An optimized imaging mass cytometry analysis framework for benchmarking segmentation and data exploration. *Cytometry Part A*, *105*(1), 36–53.

Hussein, I. H., & Raad, M. (2015). Once Upon a Microscopic Slide: The Story of Histology. *Journal of Cytology & Histology*, *06*(06). (Publisher: OMICS Publishing Group) doi: 10.4172/2157-7099.1000377

Hussein, N., Reinhard, B., Adrian, S., Marie-Lisa, E., Philipp, L., & Katarzyna, B. (2023, June). *LyNSeC: Lymphoma Nuclear Segmentation and Classification.* Zenodo. Retrieved 2024-05-22, from `https://zenodo.org/records/8065174` doi: 10.5281/zenodo.8065174

Iakubovskii, P. (2024, May). *qubvel/ttach.* Retrieved 2024-05-24, from `https://github.com/qubvel/ttach` (original-date: 2019-10-01T16:08:55Z)

Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., & Maier-Hein, K. H. (2021, February). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, *18*(2), 203–211. doi: 10.1038/s41592-020-01008-z

Israel, U., Marks, M., Dilip, R., Li, Q., Yu, C., Laubscher, E., ... others (2025). Cellsam: a foundation model for cell segmentation. *BioRxiv*, 2023–11.

Jack, N. P., Tsai, Y.-H. H., Marick, L., & Yamada, M. (2021, July). *Extention to the TNBC dataset, brain section and cell type.* Zenodo. doi: 10.5281/zenodo.3552674

Jackson, H. W., Fischer, J. R., Zanotelli, V. R. T., Ali, H. R., Mechera, R., Soysal, S. D., ... other consortium authors (2020). The single-cell pathology landscape of breast cancer. *Nature*, *578*(7796), 615–620. doi: 10.1038/s41586-019-1876-x

Jiang, S., Chan, C. N., Rovira-Clavé, X., Chen, H., Bai, Y., Zhu, B., ... Nolan, G. P. (2022, June). Combined protein and nucleic acid imaging reveals virus-dependent B cell and macrophage immunosuppression of tissue microenvironments. *Immunity*, *55*(6), 1118–1134.e8. doi: 10.1016/j.immuni.2022.03.020

Jiménez-Sánchez, D., Ariz, M., Chang, H., Matias-Guiu, X., de Andrea, C. E., & Ortiz-de Solórzano, C. (2022). NaroNet: Discovery of tumor microenvironment elements from highly multiplexed images. *Medical Image Analysis*, *78*, 102384. doi: 10.1016/j.media.2022.102384

Jones, T. R., Kang, I. H., Wheeler, D. B., Lindquist, R. A., Papallo, A., Sabatini, D. M., ... Carpenter, A. E. (2008, November). CellProfiler Analyst: data exploration and analysis software for complex image-based screens. *BMC Bioinformatics*, *9*(1), 482. doi: 10.1186/1471-2105-9-482

Kaimal, J., Thul, P., Xu, H., Ouyang, W., & Lundberg, E. (2024, June). *hpa/hpa-cell-image- segmentation-dataset.* Zenodo. Retrieved from `https://doi.org/10.5281/zenodo.13219877` doi: 10.5281/zenodo.13219877

Keren, L., Bosse, M., Marquez, D., Angoshtari, R., Jain, S., Varma, S., ... Angelo, M. (2018, September). A Structured Tumor-Immune Microenvironment in Triple Negative Breast Cancer Revealed by Multiplexed Ion Beam Imaging. *Cell*, *174*(6), 1373–1387.e19. doi: 10.1016/j.cell.2018.08.039

Khanam, R., & Hussain, M. (2024). What is yolov5: A deep look into the internal features of the popular object detector. *arXiv preprint arXiv:2407.20892*.

Kiemen, A. L., Braxton, A. M., Grahn, M. P., Han, K. S., Babu, J. M., Reichel, R., ... others (2022). Coda: quantitative 3d reconstruction of large tissues at cellular resolution. *Nature Methods*, *19*(11), 1490–1499.

Kingma, D. P., & Ba, J. (2017, January). *Adam: A Method for Stochastic Optimization.* arXiv. (arXiv:1412.6980 [cs]) doi: 10.48550/arXiv.1412.6980

Kint, S., Younes, S. T., Bao, S., Long, G., Wouters, D., Stephenson, E., ... others (2025). Spatial epigenomic niches underlie glioblastoma cell state plasticity. *bioRxiv*, 2025–05.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... Girshick, R. (2023, April). *Segment Anything.* arXiv. (arXiv:2304.02643 [cs]) doi: 10.48550/arXiv.2304.02643

Klöckner, P., Teixeira, J., Montezuma, D., Fraga, J., Horlings, H. M., Cardoso, J. S., & Oliveira, S. P. (2025, July). H&E to IHC virtual staining methods in breast cancer: an overview and benchmarking. *npj Digital Medicine*, *8*(1), 384. doi: 10.1038/s41746-025-01741-9

Koohbanani, N. A., Jahanifar, M., Tajadin, N. Z., & Rajpoot, N. (2020). Nuclick: a deep learning framework for interactive segmentation of microscopic images. *Medical Image Analysis*, *65*, 101771.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*.

Kulikov, V., & Lempitsky, V. (2020, June). Instance Segmentation of Biological Images Using Harmonic Embeddings. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3842–3850). Seattle, WA, USA: IEEE. doi: 10.1109/CVPR42600.2020.00390

Kumar, N., Verma, R., Sharma, S., Bhargava, S., Vahadane, A., & Sethi, A. (2017, July). A Dataset and a Technique for Generalized Nuclear Segmentation for Computational Pathology. *IEEE transactions on medical imaging*, *36*(7), 1550–1560. doi: 10.1109/TMI.2017.2677499

Lalit, M., Tomancak, P., & Jug, F. (2022, October). EmbedSeg: Embedding-based Instance Segmentation for Biomedical Microscopy Data. *Medical Image Analysis*, *81*, 102523. doi: 10.1016/j.media.2022.102523

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, *1*(4), 541–551.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (2002). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Lee, J., Lee, Y., Kim, J., Kosiorek, A., Choi, S., & Teh, Y. W. (2019). Set transformer: A framework for attention-based permutation-invariant neural networks. In *International conference on machine learning* (pp. 3744–3753).

Le Goc, G., d'Humières, A., Lesage, A., & Bouvier, J. (2025). cuisto: A python package to quantify neurohistological data from qupath and abba. *Journal of Open Source Software*, *10*(111), 7906.

Li, Z., Su, T., Zhang, B., Han, W., Zhang, S., Sun, G., ... Gao, X. (2024). His-mmdm: Multi-domain and multi-omics translation of histopathological images with diffusion models. *medRxiv*. doi: 10.1101/2024.07.11.24310294

Lin, J., Chen, Y.-A., Campton, D., Cooper, J., Coy, S., Yapp, C., ... Sorger, P. K. (2023, February). *High-plex immunofluorescence imaging and traditional histology of the same tissue section for discovering image-based biomarkers.* Zenodo. Retrieved from `https://doi.org/10.5281/zenodo.7637988` doi: 10.5281/zenodo.7637988

Lin, J.-R., Chen, Y.-A., Campton, D., Cooper, J., Coy, S., Yapp, C., ... Sorger, P. K. (2023). High-plex immunofluorescence imaging and traditional histology of the same tissue section for discovering image-based biomarkers. *Nature Cancer*, *4*(7), 1036–1052. doi: 10.1038/s43018-023-00576-1

Lin, J.-R., Izar, B., Wang, S., Yapp, C., Mei, S., Shah, P. M., ... Sorger, P. K. (2018, July). Highly multiplexed immunofluorescence imaging of human tissues and tumors using t-CyCIF and conventional optical microscopes. *eLife*, *7*, e31657. (Publisher: eLife Sciences Publications, Ltd) doi: 10.7554/eLife.31657

Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 2980–2988).

Lin, Z., Wei, D., Gupta, A., Liu, X., Sun, D., & Pfister, H. (2023, October). *Structure-Preserving Instance Segmentation via Skeleton-Aware Distance Transform.* arXiv. (arXiv:2310.05262 [cs]) doi: 10.48550/arXiv.2310.05262

Liu, B., Dolz, J., Galdran, A., Kobbi, R., & Ayed, I. B. (2024). Do we really need dice? the hidden region-size biases of segmentation losses. *Medical Image Analysis*, *91*, 103015.

Liu, C. C., Greenwald, N. F., Kong, A., McCaffrey, E. F., Leow, K. X., Mrdjen, D., ... Angelo, M. (2023). *Robust phenotyping of highly multiplexed tissue imaging data using pixel-level clustering.* bioRxiv. (Pages: 2022.08.16.504171 Section: New Results) doi: 10.1101/2022.08.16.504171

Liu, S., Zhang, B., Liu, Y., Han, A., Shi, H., Guan, T., & He, Y. (2021). Unpaired stain transfer using pathology-consistent constrained generative adversarial networks. *IEEE Transactions on Medical Imaging*, *40*(8), 1977-1989. doi: 10.1109/TMI.2021.3069874

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C. (2016). Ssd: Single shot multibox detector. In *Computer vision–eccv 2016: 14th european conference, amsterdam, the netherlands, october 11–14, 2016, proceedings, part i 14* (pp. 21–37).

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. *CoRR*, *abs/2103.14030*.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3431–3440).

Ma, J., Chen, J., Ng, M., Huang, R., Li, Y., Li, C., ... Martel, A. L. (2021). Loss odyssey in medical image segmentation. *Medical image analysis*, *71*, 102035.

Ma, J., Xie, R., Ayyadhury, S., Ge, C., Gupta, A., Gupta, R., ... Wang, B. (2024). The multi-modality cell segmentation challenge: Towards universal solutions. *Nature Methods*, *21*, 1103–1113. doi: https://doi.org/10.1038/s41592-024-02233-6

Mahbod, A., Polak, C., Feldmann, K., Khan, R., Gelles, K., Dorffner, G., ... Ellinger, I. (2024, January). *NuInsSeg: A fully annotated dataset for nuclei instance segmentation in H&E-stained histological images.* Zenodo. doi: 10.5281/zenodo.10518968

Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S., Stankovic, M., Scholz, P., ... others (2018). Why rankings of biomedical image analysis competitions should be interpreted with care. *Nature communications*, *9*(1), 5217.

Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M. D., Buettner, F., Christodoulou, E., ... others (2024). Metrics reloaded: recommendations for image analysis validation. *Nature methods*, *21*(2), 195–212.

McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

McNicoll, M.-M., Himdi, L., Tewfik, M., Di Maio, S., Guiot, M.-C., & Larouche, V. (2025). An unexpected case of frontal headache: Silent corticotroph pituitary neuroendocrine tumor presenting as a sphenoid sinus mass. *SAGE Open Medical Case Reports*, *13*, 2050313X251332081.

Mears, K. S., Ibrahim, K., Allen, P. M., Chinai, J. M., Avila, O. I., Muscato, A. J., ... others (2025). In vivo generation of chimeric antigen receptor t cells using optimally retargeted and functionalized lentiviral vectors with reduced immune clearance. *bioRxiv*, 2025–04.

Meijering, E. (2012, September). Cell Segmentation: 50 Years Down the Road [Life Sciences]. *IEEE Signal Processing Magazine*, *29*(5), 140–145. (Conference Name: IEEE Signal Processing Magazine) doi: 10.1109/MSP.2012.2204190

Merchant, F., & Castleman, K. (2022). *Microscope Image Processing*. Academic Press. (Google-Books-ID: IGFlEAAAQBAJ)

Midden, D., Studer, L., Hermsen, M., Farris, A., Kers, J., Hilbrands, L., & van der Laak, J. (2024). Introducing the monkey challenge: Machine-learning for optimal detection of inflammatory cells in the kidney. In *Proceedings of the european congress on digital pathology 2024.*

Milletari, F., Navab, N., & Ahmadi, S.-A. (2016). V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3d vision (3dv)* (pp. 565–571).

Naylor, P., Laé, M., Reyal, F., & Walter, T. (2018). Segmentation of nuclei in histopathology images by deep regression of the distance map. *IEEE transactions on medical imaging*, *38*(2), 448–459.

Naylor, P., Lae, M., Reyal, F., & Walter, T. (2019, February). Segmentation of Nuclei in Histopathology Images by Deep Regression of the Distance Map. *IEEE transactions on medical imaging*, *38*(2), 448–459. doi: 10.1109/TMI.2018.2865709

Ndacayisaba, L. J., Rappard, K. E., Shishido, S. N., Ruiz Velasco, C., Matsumoto, N., Navarez, R., ... others (2022). Enrichment-free single-cell detection and morphogenomic profiling of myeloma patient samples to delineate circulating rare plasma cell clones. *Current Oncology*, *29*(5), 2954–2972.

Neven, D., Brabandere, B. D., Proesmans, M., & Van Gool, L. (2019, June). Instance Segmentation by Jointly Optimizing Spatial Embeddings and Clustering Bandwidth. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 8829–8837). Long Beach, CA, USA: IEEE. doi: 10.1109/CVPR.2019.00904

O'Hurley, G., Sjöstedt, E., Rahman, A., Li, B., Kampf, C., Pontén, F., ... Lindskog, C. (2014). Garbage in, garbage out: A critical evaluation of strategies used for validation of immunohistochemical biomarkers. *Molecular Oncology*, *8*(4), 783-798. doi: https://doi.org/10.1016/j.molonc.2014.03.008

Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., ... others (2018). Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*.

Otálora, S., Marini, N., Podareanu, D., Hekster, R., Tellez, D., Van Der Laak, J., ... Atzori, M. (2022). stainlib: a python library for augmentation and normalization of histopathology h&e images. *bioRxiv*. doi: 10.1101/2022.05.17.492245

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

Ouyang, W., Beuttenmueller, F., Gómez-de Mariscal, E., Pape, C., Burke, T., Garcia-López-de Haro, C., ... others (2022). Bioimage model zoo: a community-driven resource for accessible deep learning in bioimage analysis. *BioRxiv*, 2022–06.

Pachitariu, M., Rariden, M., & Stringer, C. (2025). Cellpose-sam: superhuman generalization for cellular segmentation. *bioRxiv*, 2025–04.

Padovani, F., Mairhörmann, B., Falter-Braun, P., Lengefeld, J., & Schmoller, K. M. (2022). Segmentation, tracking and cell cycle analysis of live-cell imaging data with cell-acdc. *BMC biology*, *20*(1), 174.

Pang, M., Roy, T. K., Wu, X., & Tan, K. (2025). Cellotype: a unified model for segmentation and classification of tissue images. *Nature methods*, *22*(2), 348–357.

Pantanowitz, L., Sharma, A., Carter, A. B., Kurc, T., Sussman, A., & Saltz, J. (2018). Twenty years of digital pathology: an overview of the road travelled, what is on the horizon, and the emergence of vendor-neutral archives. *Journal of pathology informatics*, *9*(1), 40.

Parwani, A. V. (Ed.). (2022). *Whole Slide Imaging: Current Applications and Future Directions*. Cham: Springer International Publishing. doi: 10.1007/978-3-030-83332-9

Paszke, A. (2019). Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Qian, S., & Zhang, X. (2025). Characterizing tumor microenvironment heterogeneity in ebv+ ntnkl versus enktcl using spatial transcriptomics and mif. *Hematological Oncology*, *43*, e495_70096.

Qin, D., Leichner, C., Delakis, M., Fornoni, M., Luo, S., Yang, F., ... Howard, A. (2024). *MobileNetV4 – Universal Models for the Mobile Ecosystem.* arXiv. (arXiv:2404.10518 [cs]) doi: 10.48550/arXiv.2404.10518

Qiu, M., Zhou, B., Lo, F., Cook, S., Chyba, J., Quackenbush, D., ... Zhou, Y. (2020, July). A cell-level quality control workflow for high-throughput image analysis. *BMC Bioinformatics*, *21*(1), 280. doi: 10.1186/s12859-020-03603-5

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779–788).

Ren, S., He, K., Girshick, R., & Sun, J. (2016, January). *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.* arXiv. (arXiv:1506.01497 [cs]) doi: 10.48550/arXiv.1506.01497

Romera, E., Álvarez, J. M., Bergasa, L. M., & Arroyo, R. (2018, January). ERFNet: Efficient Residual Factorized ConvNet for Real-Time Semantic Segmentation. *IEEE Transactions on Intelligent Transportation Systems*, *19*(1), 263–272. (Conference Name: IEEE Transactions on Intelligent Transportation Systems) doi: 10.1109/TITS.2017.2750080

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In N. Navab, J. Hornegger, W. M. Wells, & A. F. Frangi (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (pp. 234–241). Cham: Springer International Publishing. doi: 10.1007/978-3-319-24574-4_28

Ruifrok, A. C., Johnston, D. A., et al. (2001). Quantification of histochemical staining by color deconvolution. *Analytical and quantitative cytology and histology*, *23*(4), 291–299.

Rumberger, J. L., Greenwald, N. F., Ranek, J. S., Boonrat, P., Walker, C., Franzen, J., ... Angelo, M. (2024). *Automated classification of cellular expression in multiplexed imaging data with Nimbus.* bioRxiv. (Pages: 2024.06.02.597062 Section: New Results) doi: 10.1101/2024.06.02.597062

Schapiro, D., Jackson, H. W., Raghuraman, S., Fischer, J. R., Zanotelli, V. R. T., Schulz, D., ... Bodenmiller, B. (2017, September). histoCAT: analysis of cell phenotypes and interactions in multiplex image cytometry data. *Nature Methods*, *14*(9), 873–876. doi: 10.1038/nmeth.4391

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., ... Cardona, A. (2012, July). Fiji: an open-source platform for biological-image analysis. *Nature Methods*, *9*(7), 676–682. doi: 10.1038/nmeth.2019

Schmidt, U., Weigert, M., Broaddus, C., & Myers, G. (2018). Cell Detection with Star-Convex Polygons. In A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, & G. Fichtinger (Eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018* (pp. 265–273). Cham: Springer International Publishing. doi: 10.1007/978-3-030-00934-2_30

Schneider, C. A., Rasband, W. S., & Eliceiri, K. W. (2012). Nih image to imagej: 25 years of image analysis. *Nature methods*, *9*(7), 671–675.

Shaban, M., Bai, Y., Qiu, H., Mao, S., Yeung, J., Yeo, Y. Y., ... Mahmood, F. (2024, January). MAPS: pathologist-level cell type annotation from tissue images through machine learning. *Nature Communications*, *15*(1), 28. doi: 10.1038/s41467-023-44188-w

Singer, C. (1914). Notes on the Early History of Microscopy. *Proceedings of the Royal Society of Medicine*, *7*(Sect Hist Med), 247–279. doi: 10.1177/003591571400701617

Sirinukunwattana, K., Pluim, J. P., Chen, H., Qi, X., Heng, P.-A., Guo, Y. B., ... others (2017). Gland segmentation in colon histology images: The glas challenge contest. *Medical image analysis*, *35*, 489–502.

Smith, J. A., Bloss, D. T., Williams, M. D., Posgai, A. L., Brusko, T. M., Atkinson, M. A., ... Brusko, M. A. (2025). Protocol for processing and analyzing multiplexed images improves lymphatic cell identification and spatial architecture in human tissue. *STAR Protocols*, *6*(3), 103976.

Sofroniew, N., Lambert, T., Bokota, G., Nunez-Iglesias, J., Sobolewski, P., Sweet, A., ... Zhao, R. (2025, July). *napari: a multi-dimensional image viewer for Python.* Zenodo. Retrieved from `https://doi.org/10.5281/zenodo.15779115` doi: 10.5281/zenodo.15779115

Solórzano, C., Kozubek, M., Meijering, E., & Barrutia, A. (2015). *Isbi cell tracking challenge.*

Solovyev, R., Wang, W., & Gabruseva, T. (2021). Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing*, *107*, 104117.

Stirling, D. R., Swain-Bowden, M. J., Lucas, A. M., Carpenter, A. E., Cimini, B. A., & Goodman, A. (2021). CellProfiler 4: improvements in speed, utility and usability. *BMC Bioinformatics*, *22*(1), 433. doi: 10.1186/s12859-021-04344-9

Stringer, C., & Pachitariu, M. (2024). Transformers do not outperform cellpose. *bioRxiv*, 2024–04.

Stringer, C., Wang, T., Michaelos, M., & Pachitariu, M. (2021, January). Cellpose: a generalist algorithm for cellular segmentation. *Nature Methods*, *18*(1), 100–106. doi: 10.1038/s41592-020-01018-x

Toma, M. I., Li, Y., Kehl, M., Kempchen, T. N., Esser, L., Baschun, K., ... others (2025). Low-branching vessel architecture shapes immune cell niches and predicts immune responses in renal cancer. *medRxiv*, 2025–06.

Traag, V. A., Waltman, L., & Van Eck, N. J. (2019). From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, *9*(1), 1–12.

Ulman, V., Maška, M., Magnusson, K. E. G., Ronneberger, O., Haubold, C., Harder, N., ... Ortiz-de Solorzano, C. (2017, December). An objective comparison of cell-tracking algorithms. *Nature Methods*, *14*(12), 1141–1152. doi: 10.1038/nmeth.4473

VandeLoo, A. D., Malta, N. J., Aponte, E., van Zyl, C., Xu, D., & Forest, C. R. (2025). Samcell: Generalized label-free biological cell segmentation with segment anything. *bioRxiv*.

Van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., ... Yu, T. (2014). scikit-image: image processing in python. *PeerJ*, *2*, e453.

Vanea, C., Džigurski, J., Rukins, V., Dodi, O., Siigur, S., Salumäe, L., ... Nellåker, C. (2024, March). Mapping cell-to-tissue graphs across human placenta histology whole slide images using deep learning with HAPPY. *Nature Communications*, *15*(1), 2710. doi: 10.1038/s41467-024-46986-2

van Ooij, C. (2009). Recipe for fluorescent antibodies. *Nature Cell Biology*, *11*(1), S10–S11. doi: 10.1038/ncb1932

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Vieco-Martí, I., López-Carrasco, A., Navarro, S., Granados Aparici, S., & Noguera, R. (2025). Daneelpath: Open-source digital analysis tools for histopathological research. applications in neuroblastoma models. *bioRxiv*, 2025–03.

Vázquez-García, I., Uhlitz, F., Ceglia, N., Lim, J. L. P., Wu, M., Mohibullah, N., ... Shah, S. P. (2022, December). Ovarian cancer mutational processes drive site-specific immune evasion. *Nature*, *612*(7941), 778–786. doi: 10.1038/s41586-022-05496-1

Wahlin, S. (2025). *Cellular and molecular cartography of muscle-invasive bladder cancer: Who benefits from neoadjuvant chemotherapy?* (No. 2025: 56). Lund University.

Wang, R., Qiu, Y., Hao, X., Jin, S., Gao, J., Qi, H., ... Xu, H. (2024, July). Simultaneously segmenting and classifying cell nuclei by using multi-task learning in multiplex immunohistochemical tissue microarray sections. *Biomedical Signal Processing and Control*, *93*, 106143. doi: 10.1016/j.bspc.2024.106143

Wang, X. J., Dilip, R., Bussi, Y., Brown, C., Pradhan, E., Jain, Y., ... Van Valen, D. (2024). Generalized cell phenotyping for spatial proteomics with language-informed vision models. *bioRxiv*. (preprint) doi: 10.1101/2024.11.02.621624

Weigert, M., & Schmidt, U. (2022). Nuclei instance segmentation and classification in histopathology images with stardist. In *2022 IEEE international symposium on biomedical imaging challenges (isbic)* (p. 1-4). doi: 10.1109/ISBIC56247.2022.9854534

Wightman, R. (2020). *Pytorch image models.* `https://github.com/huggingface/pytorch-image-models`. (Accessed: 2025-07-07)

Wikström, F., & Hammer, I. (2025). Evaluation of immune infiltration in a population-based cohort of diffuse large b-cell lymphoma. *LUP Student Papers*.

Windhager, J., Zanotelli, V. R. T., Schulz, D., Meyer, L., Daniel, M., Bodenmiller, B., & Eling, N. (2023, November). An end-to-end workflow for multiplexed image processing and analysis. *Nature Protocols*, *18*(11), 3565–3613. doi: 10.1038/s41596-023-00881-0

Wolf, F. A., Angerer, P., & Theis, F. J. (2018). Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, *19*(1), 15. doi: 10.1186/s13059-017-1382-0

Wolf, S., Pape, C., Bailoni, A., Rahaman, N., Kreshuk, A., Kothe, U., & Hamprecht, F. (2018). The mutex watershed: efficient, parameter-free image partitioning. In *Proceedings of the european conference on computer vision (eccv)* (pp. 546–562).

Xiao, X., Qiao, Y., Jiao, Y., Fu, N., Yang, W., Wang, L., ... Han, J. (2021, September). Dice-XMBD: Deep Learning-Based Cell Segmentation for Imaging Mass Cytometry. *Frontiers in Genetics*, *12*. (Publisher: Frontiers) doi: 10.3389/fgene.2021.721229

Yang, J., Huang, X., He, Y., Xu, J., Yang, C., Xu, G., & Ni, B. (2021). Reinventing 2d convolutions for 3d images. *IEEE Journal of Biomedical and Health Informatics*, *25*(8), 3009–3018.

Yu, W., Lee, H. K., Hariharan, S., Bu, W. Y., & Ahmed, S. (2025). *CCDB:6843, Mus musculus Neuroblastoma cell and nuclear segmentation dataset.* Cell Image Library dataset. Retrieved from `https://www.cellimagelibrary.org/images/CCDB_6843` (Accessed 31 July 2025) doi: 10.7295/W9CCDB6843

Zaheer, M., Kottur, S., Ravanbhakhsh, S., Póczos, B., Salakhutdinov, R., & Smola, A. J. (2017, December). Deep Sets. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 3394–3404). Red Hook, NY, USA: Curran Associates Inc.

Zheng, Y., Abila, E., Chrenková, E., Winkler, J., & Rendeiro, A. F. (2025). Lazyslide: accessible and interoperable whole slide image analysis. *BioRxiv*, 2025–05.

Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., & Liang, J. (2019). Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE transactions on medical imaging*, *39*(6), 1856–1867.

Zhu, B., Gao, S., Chen, S., Yeung, J., Bai, Y., Huang, A. Y., ... others (2024). Cross-domain information fusion for enhanced cell population delineation in single-cell spatial-omics data. *bioRxiv*.