

FROM SPECIFIC TO UNIVERSAL: ONE BIOMEDICAL IMAGE SEGMENTATION MODEL TO RULE THEM ALL

by

SEYED ALIREZA VAEZI

(Under the Direction of Shannon Quinn)

ABSTRACT

Biomedical image segmentation is essential for clinical diagnostics, treatment planning, and fundamental biological research. Yet, the diversity of imaging modalities, complex structures, and the limited availability of annotated datasets pose significant challenges. Although convolutional neural networks (CNNs) often deliver state-of-the-art performance, they typically require extensive labeled data and domain-specific adjustments, restricting their generalizability. To address these limitations, self-supervised and unsupervised segmentation methods have emerged, exploiting unlabeled or weakly labeled data to alleviate the annotation burden. However, these methods can struggle with segmentation accuracy and robustness when confronted with complex domain-specific variability. Recent advances in foundation models, particularly the Segment Anything Model (SAM), suggest a promising path forward. Trained on large-scale, diverse data, SAM enables generalized segmentation via zero-shot learning, indicating potential applicability across a range of biomedical imaging domains. Nonetheless, fine-tuning and adaptation are necessary to ensure reliable performance and reproducibility in specialized biomedical contexts. This dissertation bridges these critical gaps. This dissertation introduces a supervised pipeline for 3D cell instance segmentation, tracking, and motility classification, centering on *Toxoplasma gondii*, alongside self-supervised and minimally supervised approaches that reduce the annotation burden and enhance model generalization across diverse imaging contexts. Additionally, the adaptation of foundation models such as SAM is explored, detailing fine-tuning techniques that empower reliable biomedical segmentation in specialized applications. Collectively, these contributions advance the field of biomedical imaging by mitigating annotation requirements, improving robustness, and widening applicability. By synthesizing supervised, self-supervised, and foundation model-based strategies, this dissertation offers a cohesive framework that addresses critical hurdles including annotation burden, domain generalization, and model adaptability, paving the way for more efficient, scalable, and broadly accessible biomedical segmentation solutions.

INDEX WORDS: [Biomedical Image Segmentation, Supervised Learning, Unsupervised Learning, Foundation Models, Segment Anything Model, 3D Cell Segmentation, Cell Tracking, Cilia Segmentation, Domain Generalization, *Toxoplasma gondii*]

FROM SPECIFIC TO UNIVERSAL: ONE BIOMEDICAL IMAGE SEGMENTATION MODEL TO
RULE THEM ALL

by

SEYED ALIREZA VAEZI

M.S., Iran University of Science and Technology (IUST), Iran, 2016

A Dissertation Submitted to the Graduate Faculty of the
University of Georgia in Partial Fulfillment of the Requirements for the
Degree.

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2025

FROM SPECIFIC TO UNIVERSAL: ONE BIOMEDICAL IMAGE SEGMENTATION MODEL TO
RULE THEM ALL

by

SEYED ALIREZA VAEZI

Major Professor: Shannon Quinn

Committee: Hamid R. Arabnia
Tianming Liu
Kyle Johnsen

Electronic Version Approved:

Ron Walcott
Dean of the Graduate School
The University of Georgia
May 2025

DEDICATION

To my parents, Ali and Azam, whose unconditional love, sacrifice, and wisdom taught me resilience and compassion, guiding me every step of the way. And to Niloofar, my partner and best friend, whose boundless energy, unwavering belief in me, and remarkable strength showed me what perseverance and passion truly mean. Your presence in my life has made every challenge surmountable and every victory sweeter.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my advisor, Dr. Shannon Quinn, whose endless support, genuine care, and remarkable patience guided me through this Ph.D. journey. Thank you for standing by me and patiently helping me find direction whenever I felt most lost. Your encouragement, understanding, and unwavering belief in my capabilities have shaped not only my research but also my growth as an individual and scholar.

CONTENTS

Acknowledgments	v
List of Figures	viii
List of Tables	x
1 Introduction	1
1.1 In Search of a Universal Biomedical Segmentation Model	1
1.2 Challenges	3
2 Deep Models For Supervised Image Segmentation	7
2.1 Introduction	7
2.2 Background	8
2.3 Methodologies	10
2.4 Evaluation	14
2.5 Conclusion and Final Remarks	16
2.6 Limitations and Future Directions	18
3 Training a Supervised Cilia Segmentation Model from Self-Supervision	21
3.1 Introduction	21
3.2 Background	22
3.3 Methodology	23
3.4 Results and Discussion	27
3.5 Conclusions and Final Remarks	29
4 Minimally-Supervised Biomedical image Segmentation via Contrastive Learning	32
4.1 Introduction	32
4.2 Background	33
4.3 Methodology	34
4.4 Results and Discussion	37
4.5 Conclusion and Final Remarks	41

5	Toward a Foundation Model for Biomedical Image Segmentation	42
5.1	Introduction	42
5.2	Background	43
5.3	Methodology	45
5.4	Results and Discussion	49
5.5	Conclusion and Final Remarks	54
6	Conclusion	57
6.1	Summary of Contributions	57
6.2	Theoretical and Practical Implications	58
6.3	Limitations, Failure Modes, and Future Directions	58
6.4	Broader Impact and Scientific Contributions	61
	Appendices	62
A	Appendices	62
	Bibliography	72

LIST OF FIGURES

2.1	TSeg's Napari Plugin Interface	11
2.2	On the left, a sample frame of the 3D video of <i>T. gondii</i> cells. The image is captured using a PlanApo 20x objective (NA = 0.75) on a preheated Nikon Eclipse TE300 epifluorescence microscope. On the right, the same frame after denoising.	11
2.3	The pre-processing widget includes adaptive thresholding, normalization, and noise removal to enhance image quality.	12
2.4	The CNN Detection widget integrates PlantSeg for tissue-specific 3D segmentation and CellPose for diverse cell types. These tools are implemented in the backend via their APIs, ensuring seamless operation.	13
2.5	Left, 3D connected component labeling (CCL) is used to extract features from the segmented images. Middle, the centroids of the features are calculated using the center of mass function in scipy. Right, the tracking algorithm connects centroids across time instances to track the cells.	14
2.6	The tracking widget allows the user to set the parameters for the tracking and clustering algorithms and visualizes the results.	15
2.7	Clustering of <i>T. gondii</i> motility patterns in 3D space using an autoregressive model (AR) as introduced by Fazli et al. [21]. The AR model addresses the limitations of K-means by considering geodesic distances and non-isotropic clusters.	16
3.1	A sample of three videos in our cilia dataset with their manually annotated ground truth masks.	24
3.2	Representation of rotation (curl) component of OF at a random time.	25
3.3	The pixel representation of the 5-order AR model of the OF component of a sample video. The x and y axes correspond to the width and height of the video.	26
3.4	The process of computing the masks. a) Subtracting the second-order AR parameter from the first-order, followed by b) Adaptive thresholding, which suffers from under/over-segmentation. c) A Gaussian blur filter, followed by d) An Otsu thresholding eliminates the under/over-segmentation.	26
3.5	The model predictions on 5 dyskinetic cilia samples. The first column shows a frame of the video, the second column shows the manually labeled ground truth, the third column is the model's prediction, and the last column is a thresholded version of the prediction.	27

4.1	(a) The architecture of our contrastive network applied to two consecutive frames. (b) The internals of an MBConv layer.	35
4.2	Visual comparison of segmentation performance across different datasets. The variation in performance across datasets indicates the challenges caused by different imaging modalities and cell types.	38
5.1	BiomedCLIP’s out-of-the-box performance over CTC’s 2D datasets	47
5.2	Performance of off-the-shelf BiomedCLIP with Short, Medium, and Detailed prompts, demonstrating poor localization without fine-tuning.	50
5.3	Localization results after fine-tuning BiomedCLIP for 32 epochs on full images without masks, showing incorrect and overly specialized localization.	51
5.4	Impact of adjusted hardness parameters ($\beta_1 = 0.65$, $\beta_2 = 0.65$) during fine-tuning, showing improved localization diversity but potentially higher false-positive rates compared to baseline fine-tuning.	52
A.1	Improved ciliary region localization after fine-tuning BiomedCLIP for one epoch using masked images with short captions.	63
A.2	Effective and balanced localization achieved by fine-tuning BiomedCLIP with a reduced dataset size (small train) using masked images and short captions for 5 epochs.	64
A.3	Localization results demonstrating increased flexibility but higher false positives when fine-tuning BiomedCLIP using long captions.	65
A.4	Localization impact of higher temperature ($\tau = 0.4$) during fine-tuning, resulting in blurred similarity distributions and compromised accuracy.	66

LIST OF TABLES

2.1	Top-Performing Jaccard Index Scores on 2D CTC Datasets.	17
2.2	Best Segmentation Performance on 2D CTC Datasets (Jaccard Index)	18
2.3	Best Segmentation Performance on 3D CTC Datasets (Jaccard Index)	18
3.1	Summary of model architecture, training setup, and dataset distribution	28
3.2	The performance of the model in validation and testing phases.	29
3.3	Comparison of Cilia Segmentation Performance Metrics with Zain et al. [109].	29
4.1	Cell Tracking Challenge (CTC) 2D Datasets	36
4.2	Dice coefficients and intersection-over-union (IoU) scores for the CTC datasets.	39
4.3	Top-Performing Jaccard Index Scores on 2D CTC Datasets.	39
5.1	Fine-Tuning Parameter Variations	47
5.2	Effects of parameter changes on DHN-NCE loss.	49
5.3	Average IoU and Dice scores of BiomedCLIP models under different training configurations	49
5.4	Average Per-Pixel Confusion Matrix Metrics (15 Test Samples)	53
5.5	Caption Generation Logic Based on Image Content for Fine-Tuning.	53
A.1	Cell Tracking Challenge Datasets	67
A.2	Overview of Cell Tracking Challenge Datasets with Sample Images.	68
A.3	CellPose Performance on 2D Datasets	69
A.4	PlantSeg Performance on 2D Datasets	70
A.5	CellPose Performance on 3D Datasets	70
A.6	PlantSeg Performance on 3D Datasets	71

CHAPTER I

INTRODUCTION

I.1 In Search of a Universal Biomedical Segmentation Model

Biomedical image segmentation is critical for accurate diagnosis, treatment planning, and biological research, as it precisely delineates anatomical structures and pathological changes. Accurate segmentation enables quantitative analyses, enhancing objectivity, consistency, and reproducibility in medical assessments. Historically, image segmentation relied on traditional methods such as thresholding, watershed segmentation, and optical flow, primarily utilizing pixel intensity to differentiate image regions. However, these traditional techniques frequently encountered significant difficulties when applied to noisy imaging conditions, complex anatomical structures, or subtle intensity gradients, severely limiting their practical applicability in clinical and research settings. Addressing these limitations necessitated sophisticated computational approaches, driving substantial methodological evolution toward advanced image segmentation strategies [26], [38].

Traditional rule-based segmentation methods, including thresholding and watershed algorithms, operate by direct pixel intensity analysis to define boundaries but exhibit limited robustness in challenging imaging conditions characterized by noise or poor contrast. Machine learning-based segmentation techniques, such as Support Vector Machines (SVMs) [14], Random Forest classifiers [4], and contrastive learning methods [10], subsequently enhanced segmentation capabilities through statistical pattern recognition. These methods provided improved flexibility and accuracy by learning complex data patterns from annotated examples. Despite these advancements, traditional machine learning methods required intensive manual feature engineering, constraining scalability and limiting generalization across diverse biomedical datasets and modalities.

The advent of deep learning approaches effectively addressed these limitations, signifying a transformative shift in image segmentation research. Deep learning-based segmentation methods, particularly those leveraging Convolutional Neural Networks (CNNs), revolutionized the field by automatically learning hierarchical feature representations directly from raw images, thus eliminating the need for manual feature extraction. CNNs inherently learn both low-level spatial features and high-level semantic features, significantly enhancing segmentation accuracy, consistency, and robustness across diverse biomedical datasets.

These deep learning-based segmentation techniques typically fall into supervised, semi-supervised, and unsupervised categories, differentiated by the amount and type of annotated data required. Among supervised methods, the U-Net architecture [81], specifically developed for biomedical image segmentation, has emerged as a cornerstone due to its highly effective feature extraction and superior generalization capabilities [9], setting foundational benchmarks for current segmentation research.

U-Net and its variants have achieved widespread adoption in biomedical image segmentation due to their exceptional ability to automate feature extraction directly from images without manual intervention or extensive preprocessing. Variants include foundational architectures such as the original U-Net [81] and 3D U-Net [13], specialized for volumetric data segmentation in computed tomography (CT) and magnetic resonance imaging (MRI). Advanced variants introduce architectural enhancements [6], including Attention U-Net for selective feature refinement [73], Inception U-Net for multi-scale context capture [112], Residual U-Net for deeper training capabilities [1], and Dense U-Net promoting extensive feature reuse to enhance accuracy and training efficiency [24]. Collectively, these CNN-based architectures currently represent the state-of-the-art in biomedical image segmentation, significantly advancing diagnostic precision and enabling more reliable quantitative analyses in medical and biological research.

Biomedical images span diverse imaging modalities, each presenting unique segmentation challenges due to varying image characteristics, resolution, noise levels, and structural complexity. Clinical imaging modalities, including CT, MRI, positron emission tomography (PET), and ultrasound imaging, each feature distinct properties affecting segmentation, such as variable tissue contrast, noise artifacts, and differing spatial resolutions [34]. Microscopy imaging modalities, such as fluorescence microscopy, bright-field microscopy, phase-contrast microscopy, and electron microscopy, present unique challenges, including intricate subcellular structures, variable staining methods, and diverse cellular morphologies. Given the extensive biological diversity, segmentation targets range widely from microscopic features like nuclei, mitochondria, and cilia to macroscopic anatomical structures like tumors, lesions, blood vessels, bones, brain anatomy. This combination of biological variability and modality-specific imaging challenges underscores the necessity for specialized yet adaptable deep learning models [38] capable of efficiently generalizing across multiple biomedical segmentation tasks without extensive reconfiguration.

Recently, universal segmentation frameworks, exemplified by the Segment Anything Model (SAM) [49], have emerged as significant advancements aiming to generalize segmentation tasks across diverse imaging domains. SAM leverages flexible user inputs, including single points, bounding boxes, and textual descriptions, enabling robust segmentation performance without requiring specialized retraining or extensive annotated datasets. Its effectiveness arises from a transformer-based architecture combined with extensive pre-training on large-scale, diverse image datasets, facilitating segmentation even of previously unseen biomedical structures and modalities, thus significantly reducing the practical barriers for broad biomedical adoption [65].

The central objective of this dissertation is to systematically address existing gaps in biomedical image segmentation methodologies, explicitly focusing on enhancing model generalizability, reducing dependency on annotated datasets, and improving usability for practical biomedical applications. Specifically, the research investigates three interconnected questions:

1. How can supervised segmentation methods be optimized to enhance performance and generalization across diverse biomedical datasets?
2. In what ways can unsupervised and self-supervised learning techniques effectively reduce the need for extensive annotated biomedical datasets without compromising segmentation accuracy?
3. How can universal segmentation frameworks be adapted and validated for robust performance in diverse biomedical imaging scenarios?

Despite these methodological advancements, significant barriers persist, hindering wide-scale adoption of deep learning segmentation methods in biomedical imaging. These ongoing challenges include limited model generalizability, heavy reliance on extensive annotated datasets [97], poor transferability across distinct biomedical imaging modalities, and practical usability constraints, especially for users lacking extensive computational expertise. The following "Challenges" section delves deeper into these issues, setting the stage for subsequent chapters, which systematically propose, develop, and evaluate novel solutions addressing these critical barriers.

1.2 Challenges

Biomedical image segmentation, particularly with deep learning models, has seen significant advancements; however, numerous practical and theoretical challenges remain, hindering wider clinical and research adoption. While architectures like nnU-Net [38] and vision foundation models have advanced biomedical segmentation, persistent challenges in cross-domain robustness and clinical translation continue to limit adoption, as evidenced by recent multi-modal validation studies [39]. The following subsections explicitly outline these challenges and the potential strategies currently employed to overcome them.

1.2.1 Generalizability and Transferability

Despite significant success, CNN-based segmentation models often exhibit poor generalizability and limited transferability across distinct biomedical imaging domains due to variability in imaging protocols, acquisition settings, and biological structures. CNN architectures like U-Net variants show excellent intra-domain performance but suffer catastrophic failure when applied cross-domain, as shown in multi-center MRI studies [17]. These limitations arise largely because obtaining large, accurately labeled biomedical datasets is costly and time-intensive, often requiring substantial expert knowledge. As a consequence, researchers often design specialized deep learning models tailored specifically to individual imaging modalities or biological structures, limiting their broader application.

Generalizability specifically denotes a model's capability to maintain robust performance on previously unseen datasets, distinct from the original training domain. This property is especially crucial in biomedical contexts, where significant variability arises due to diverse imaging protocols, acquisition devices, tissue characteristics, and patient-specific differences. Both data augmentation and transfer learning

are widely utilized strategies to mitigate overfitting and significantly enhance generalization when training datasets are limited. Despite its proven effectiveness in limited-data scenarios, transfer learning introduces specific challenges, including domain discrepancies between source and target datasets, increased computational complexity, potential biases inherited from source domains, and persistent risks of overfitting.

Chapter 5 addresses these limitations by evaluating the SAM [49] as a universal framework to bridge domain gaps, demonstrating how fine-tuning with biomedical-specific prompts enhances generalizability across modalities like MRI and fluorescence microscopy.

1.2.2 Data Scarcity and Annotation Strategies

In response to data scarcity and high annotation costs in biomedical segmentation tasks, multiple innovative solutions have emerged to leverage limited annotated data effectively. Semi-supervised learning [95], for instance, combines limited labeled datasets with abundant unlabeled data, effectively enhancing segmentation accuracy by exploiting underlying data distributions. Active learning [86] strategically queries highly informative data points for expert annotation, optimizing labeling efforts and maximizing the value of limited annotated datasets. Data augmentation approaches, including geometric image transformations and synthetic image generation via Generative Adversarial Networks (GANs) [22], effectively increase data diversity, reduce overfitting, and substantially enhance model robustness.

Transfer learning leverages knowledge gained from previously trained models on related tasks, substantially reducing the annotation burden for new segmentation applications. Self-supervised learning frameworks generate their supervisory signals via pretext tasks, such as predicting spatial positioning or reconstructing missing image regions [10], thereby facilitating robust representation learning without explicit annotations. Emerging approaches, including few-shot, one-shot, and zero-shot learning [114], specifically address extreme annotation scarcity by leveraging generalized representations, semantic meta-data, or minimal annotated examples to effectively segment unseen biomedical structures.

Chapters 3 and 4 directly address these limitations: Chapter 3 introduces self-supervised pseudo-labels derived from optical flow to segment cilia without manual masks, while Chapter 4 leverages contrastive learning to reduce annotation dependence on Cell Tracking Challenge datasets [68].

1.2.3 Complementary Role of Unsupervised Methods

Unsupervised segmentation methods, which operate without annotated datasets or extensive pre-training, typically employ domain-specific heuristics or clustering techniques to delineate structures. Although these methods generally yield lower accuracy compared to supervised CNN-based methods, they offer significant advantages in reproducibility and generalizability due to their independence from prior data annotations and their minimal reliance on task-specific training [40]. Consequently, these unsupervised approaches serve as complementary solutions, particularly beneficial in exploratory research contexts or resource-constrained scenarios.

Chapter 3 exemplifies this approach, using unsupervised motion analysis to generate cilia segmentation masks, enabling reproducible phenotyping of dyskinetic ciliary motion without labeled training data [93].

1.2.4 Reproducibility of Deep Learning Models

Reproducibility, the ability to consistently obtain similar outcomes from identical data and computational procedures, remains a critical requirement in biomedical research to validate findings and ensure reliable clinical translation. Ensuring reproducibility is fundamental to scientific integrity, enabling independent verification of results, building trust in methodologies, and facilitating reliable clinical and research translations. Reproducibility is heavily influenced by factors such as dataset variability, model architecture choices, optimization strategies, random initialization, and computational environments, each contributing potential sources of variability in outcomes [67].

Lack of reproducibility not only undermines scientific validity but also results in significant resource waste, impeded scientific progress, misleading conclusions, and ethical implications, particularly critical in sensitive biomedical applications. Recent literature emphasizes comprehensive documentation, standardized model evaluation practices, fixed random seeds, robust cross-validation frameworks, explicit reporting of evaluation metrics, and open-source sharing of code and datasets to enhance reproducibility in biomedical deep learning research [80].

To promote reproducibility, this work open-sources code for TSeg in chapter 2 and contrastive learning pipelines in chapter 4, adopts fixed random seeds, and reports metrics like Dice scores with cross-validation across all experiments in chapters 2, 3, 4, and 5, and shares relevant datasets publicly through Zenodo.

1.2.5 Practical Usability and Accessibility

While deep learning models exhibit strong performance across biomedical segmentation tasks, their practical usability remains contingent on multiple factors, including task complexity, data availability, and required model customization. For biomedical researchers and clinical practitioners with limited computational expertise, simplicity and ease of use become critical factors influencing adoption and effective utilization of deep learning-based segmentation tools [97]. These users particularly benefit from segmentation methodologies that offer robust, out-of-the-box performance without extensive hyperparameter tuning or specialized configuration.

Developing intuitive graphical user interfaces (GUIs) and interactive tools substantially improves accessibility for non-expert users, such as biologists and clinical practitioners, who require practical segmentation solutions without extensive technical knowledge. Facilitating easier interaction with deep learning models through user-friendly software not only enhances usability but also accelerates adoption and effective integration into routine clinical practice and biomedical research workflows [87].

The dissertation is organized to comprehensively address these research questions through a structured approach, encompassing supervised, self-supervised, and foundation-model-driven segmentation techniques. The chapters are organized as follows:

- **Chapter 2** introduces TSeg, a novel supervised segmentation pipeline explicitly designed for 3D cell instance segmentation, tracking, and motility classification, particularly optimized for noisy

microscopy data such as those of *Toxoplasma Gondii*. This chapter directly addresses challenges of robust supervised segmentation and practical usability, showcasing TSeg’s effectiveness and generalizability to diverse cell types through user-friendly interfaces and optimized workflows.

- **Chapter 3** explores self-supervised segmentation methods, proposing an innovative pseudo-labeling strategy to significantly reduce annotation dependency. Utilizing motion-derived masks to generate pseudo-labels, this method addresses the scarcity of annotated datasets, particularly demonstrated in challenging cilia segmentation tasks where annotation efforts are costly and complex.
- **Chapter 4** investigates minimally supervised segmentation approaches leveraging contrastive learning frameworks. By training on minimal labeled data, this chapter assesses and demonstrates the capability of contrastive learning to improve generalization and robustness across diverse biomedical imaging modalities, explicitly addressing the challenges of data scarcity and limited generalizability.
- **Chapter 5** focuses on adapting and validating universal segmentation frameworks, specifically examining the Segment Anything Model (SAM) and its applicability to biomedical contexts. Through prompt engineering and strategic fine-tuning, this chapter systematically addresses domain-specific adaptation challenges and evaluates the robustness, effectiveness, and reproducibility of SAM-based segmentation across varied biomedical imaging tasks.
- Finally, **Chapter 6** summarizes the methodological advancements described in previous chapters, discusses their broader implications for biomedical imaging research and clinical applications, identifies persistent gaps, and proposes promising directions for future research in biomedical image segmentation.

CHAPTER 2

DEEP MODELS FOR SUPERVISED IMAGE SEGMENTATION

2.1 Introduction

Quantitative cell research often requires the measurement of different cell properties including size, shape, and motility. This step is facilitated using segmentation of imaged cells. With fluorescent markers, computational tools can be used to complete segmentation and identify cell features and positions over time. 2D measurements of cells can be useful, but the more difficult task of deriving 3D information from cell images is vital for metrics such as motility and volumetric qualities.

Most of the state-of-the-art pipelines are restricted to 2D space which is not a true representative of the actual motion of the organism. Many of them require knowledge and expertise in programming, or in machine learning and deep learning models and frameworks, thus limiting the demographic of users that can use them. All of them solely include a subset of the aforementioned modules (i.e. detection, segmentation, tracking, and classification) [89]. Many pipelines rely on the user to train their own model, hand-tailored for their specific application. This demands high levels of experience and skill in ML/DL and consequently undermines the possibility and feasibility of quickly utilizing an off-the-shelf pipeline and still getting good results. PlantSeg uses a variant of 3D U-Net, called Residual 3D U-Net, for preprocessing and segmentation of multiple cell types [101]. PlantSeg performs best among Deep Learning algorithms for 3D Instance Segmentation and is very robust against image noise [42]. The segmentation module also includes the optional use of CellPose [89]. CellPose is a generalized segmentation algorithm trained on a wide range of cell types and is the first step toward increased optionality in TSeg. The Cell Tracking module consolidates the cell particles across the z-axis to materialize cells in 3D space and estimates centroids for each cell. The tracking module is also responsible for extracting the trajectories of cells based on the movements of centroids throughout consecutive video frames, which is eventually the input of the motion classifier module.

Toxoplasmosis is an infection caused by the intracellular parasite *Toxoplasma gondii*. (*T. gondii*) is one of the most successful parasites, infecting at least one-third of the world's population. Although

Toxoplasmosis is generally benign in healthy individuals, the infection has fatal implications in fetuses and immunocompromised individuals [83]. *T. gondii*'s virulence is directly linked to its lytic cycle which is comprised of invasion, replication, egress, and motility. Studying the motility of *T. gondii* is crucial in understanding its lytic cycle in order to develop potential treatments.

To address these we present TSeg. It segments *T. gondii* cells in 3D microscopic images, tracks their trajectories, and classifies the motion patterns observed throughout the 3D frames. TSeg is comprised of four modules: pre-processing, segmentation, tracking, and classification. We developed TSeg as a plugin for Napari [88] - an open-source fast and interactive image viewer for Python designed for browsing, annotating, and analyzing large multi-dimensional images. Having TSeg implemented as a part of Napari not only provides a user-friendly design but also gives more advanced users the possibility to attach and execute their custom code and even interact with the steps of the pipeline if needed. The preprocessing module is equipped with basic and extra filters and functionalities to aid in the preparation of the input data. TSeg gives its users the advantage of utilizing the functionalities that PlantSeg and CellPose provide. These functionalities can be chosen in the pre-processing, detection, and segmentation steps. This brings forth a huge variety of algorithms and pre-built models to select from, making TSeg not only a great fit for *T. gondii*, but also a variety of different cell types.

2.2 Background

The recent solutions in generalized and automated segmentation tools are focused on 2D cell images. Segmentation of cellular structures in 2D is important but not representative of realistic environments. Microbiological organisms are free to move on the z-axis and tracking without taking this factor into account cannot guarantee a full representation of the actual motility patterns. As an example, Fazli et al. [19] identified three distinct motility types for *T. gondii* with two-dimensional data, however, they also acknowledge and state that based established heuristics from previous works there are more than three motility phenotypes for *T. gondii*. The focus on 2D research is understandable due to several factors. 3D data is difficult to capture as tools for capturing 3D slices and the computational requirements for analyzing this data are not available in most research labs. Most segmentation tools are unable to track objects in 3D space as the assignment of related centroids is more difficult. The additional noise from capture and focus increases the probability of incorrect assignment. 3D data also has issues with overlapping features and increased computation required per frame of time.

Fazli et al. [19] studies the motility patterns of *T. gondii* and provides a computational pipeline for identifying motility phenotypes of *T. gondii* in an unsupervised, data-driven way. In that work Ca^{2+} is added to *T. gondii* cells inside a Fetal Bovine Serum. *T. gondii* cells react to Ca^{2+} and become motile and fluorescent. The images of motile *T. gondii* cells were captured using an LSM 710 confocal microscope. They use Python 3 and associated scientific computing libraries (NumPy, SciPy, scikit-learn, matplotlib) in their pipeline to track and cluster the trajectories of *T. gondii*. Based on this work Fazli et al. [20] work on another pipeline consisting of preprocessing, sparsification, cell detection, and cell tracking modules to track *T. gondii* in 3D video microscopy where each frame of the video consists of image slices

taken 1 micro-meters of focal depth apart along the z-axis direction. In their latest work Fazli et al. [21] developed a lightweight and scalable pipeline using task distribution and parallelism. Their pipeline consists of multiple modules: reprocessing, sparsification, cell detection, cell tracking, trajectories extraction, parametrization of the trajectories, and clustering. They could classify three distinct motion patterns in *T. gondii* using the same data from their previous work.

While combining open-source tools is not a novel architecture, little has been done to integrate 3D cell tracking tools. Fazeli et al. [18], motivated by the interest in providing robust yet accessible tools for researchers without programming expertise, developed a comprehensive pipeline combining StarDist [98] and TrackMate [91] for automated 2D cell tracking. This pipeline leverages the ZeroCostDL4Mic [7] platform, enabling researchers with no coding experience to train deep learning models on their own data, significantly lowering the barrier to entry. StarDist facilitates segmentation with star-convex polygon approximation, robustly distinguishing cells from the background even in challenging imaging conditions. TrackMate then uses these segmentation outputs to reliably track cells across timeframes, providing quantitative analytics such as velocity and trajectory characteristics. Despite its utility, the pipeline remains limited to 2D analysis, highlighting the need for extending such integrative approaches to 3D microscopy, as we propose with TSeg.

This Stardist pipeline is similar in concept to TSeg. Both create an automated segmentation and tracking pipeline but TSeg is oriented to 3D data. Cells move in 3-dimensional space that is not represented in a flat plane. TSeg also does not require the manual training necessary for the other pipeline. Individuals with low technical expertise should not be expected to create masks for training or even understand the training of deep neural networks. Lastly, this pipeline does not account for imperfect datasets without the need for preprocessing. All implemented algorithms in TSeg account for microscopy images with some amount of noise.

Wen et al. [99] combines multiple existing new technologies including deep learning and presents 3DeeCellTracker. 3DeeCellTracker segments and tracks cells on 3D time-lapse images. Using a small subset of their dataset they train the deep learning architecture 3D U-Net for segmentation. For tracking, a combination of two strategies was used to increase accuracy: local cell region strategies, and spatial pattern strategy. Kapoor et al. [41] presents VollSeg that uses deep learning methods to segment, track, and analyze cells in 3D with irregular shape and intensity distribution. It is a Jupyter Notebook-based Python package and also has a UI in Napari. For tracking, a custom tracking code is developed based on Trackmate.

Many segmentation tools require some amount of knowledge in Machine or Deep Learning concepts. Training the neural network in creating masks is a common step for open-source segmentation tools. Automating this process makes the pipeline more accessible to microbiology researchers.

2.3 Methodologies

2.3.1 Data

Our dataset consists of 11 videos of *T. gondii* cells under a microscope, obtained from different experiments with different numbers of cells. The videos are on average around 63 frames in length. Each frame has a stack of 41 image slices of size 500×502 pixels along the z-axis (z-slices). The z-slices are captured $1 \mu\text{m}$ apart in optical focal length making them $402 \mu\text{m} \times 401 \mu\text{m} \times 40 \mu\text{m}$ in volume. The slices were recorded in raw format as RGB TIF images but are converted to grayscale for our purpose. This data is captured using a PlanApo 20x objective (NA = 0.75) on a preheated Nikon Eclipse TE300 epifluorescence microscope. The image stacks were captured using an iXon 885 EMCCD camera (Andor Technology, Belfast, Ireland) cooled to -70°C and driven by NIS Elements software (Nikon Instruments, Melville, NY) as part of related research by Ward et al. [57]. The camera was set to frame transfer sensor mode, with a vertical pixel shift speed of $1.0 \mu\text{s}$, vertical clock voltage amplitude of +1, readout speed of 35MHz, conversion gain of $3.8\times$, EM gain setting of 3 and 2×2 binning, and the z-slices were imaged with an exposure time of 16ms.

2.3.2 Software

Napari Plugin

TSeg is developed as a plugin for Napari - a fast and interactive multi-dimensional image viewer for Python that allows volumetric viewing of 3D images [88]. Plugins enable developers to customize and extend the functionality of Napari. For every module of TSeg, we developed its corresponding widget in the GUI, plus a widget for file management. The widgets have self-explanatory interface elements with tooltips to guide the inexperienced user to traverse through the pipeline with ease. Layers in Napari are the basic viewable objects that can be shown in the Napari viewer. Seven different layer types are supported in Napari: Image, Labels, Points, Shapes, Surface, Tracks, and Vectors, each of which corresponds to a different data type, visualization, and interactivity [88]. After its execution, the viewable output of each widget gets added to the layers. This allows the user to evaluate and modify the parameters of the widget to get the best results before continuing to the next widget. Napari supports bidirectional communication between the viewer and the Python kernel and has a built-in console that allows users to control all the features of the viewer programmatically. This adds more flexibility and customizability to TSeg for the advanced user. The full code of TSeg is available on GitHub under the MIT open source license at <https://github.com/salirezav/tseg>. TSeg can be installed through Napari's plugins menu.

Computational Pipeline

Pre-Processing

Due to the fast imaging speed in data acquisition, the image slices will inherently have a vignetting artifact, meaning that the corners of the images will be slightly darker than the center of the image - Figure 2.2. To eliminate this artifact we added adaptive thresholding and logarithmic correction to the pre-processing

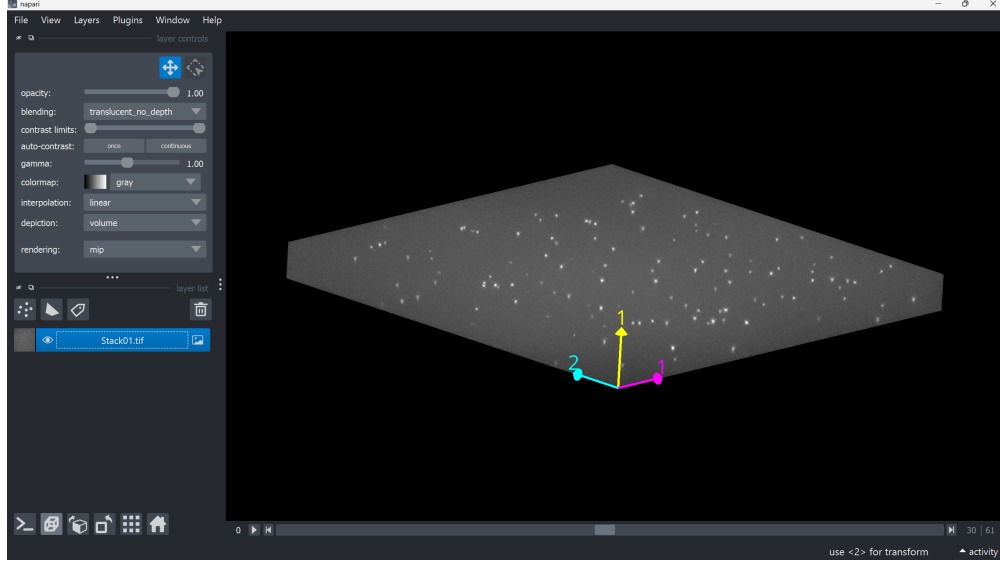


Figure 2.1: TSeg’s Napari Plugin Interface

module. Furthermore, another prevalent artifact on our dataset images was a Film-Grain noise (AKA salt and pepper noise). To remove or reduce such noise a simple gaussian blur filter and a sharpening filter are included.

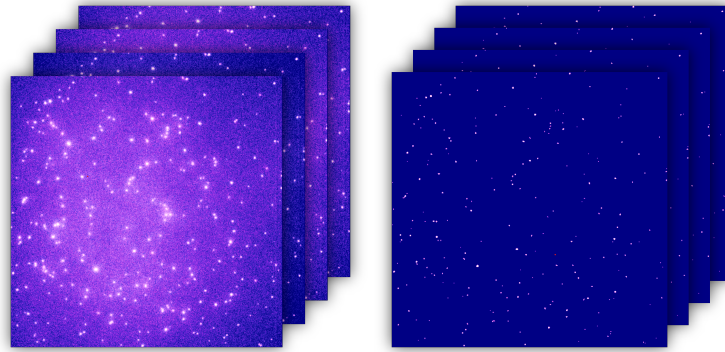


Figure 2.2: On the left, a sample frame of the 3D video of *T. gondii* cells. The image is captured using a PlanApo 20x objective (NA = 0.75) on a preheated Nikon Eclipse TE300 epifluorescence microscope. On the right, the same frame after denoising.

Cell Detection and Segmentation

TSeg’s Detection and Segmentation modules are in fact backed by PlantSeg and CellPose. The Detection Module is built only based on PlantSeg’s CNN Detection Module [101], and for the Segmentation Module, only one of the two tools can be selected to be executed as the segmentation tool in the pipeline. Naturally,

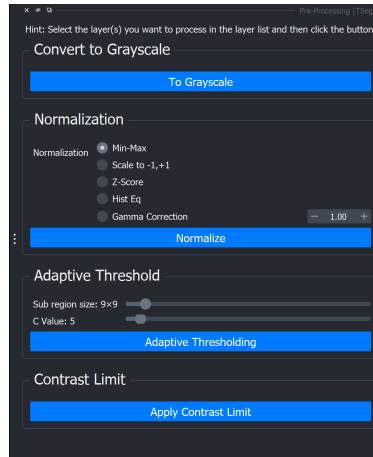


Figure 2.3: The pre-processing widget includes adaptive thresholding, normalization, and noise removal to enhance image quality.

each of the tools demands specific interface elements different from the others since each accepts different input values and various parameters. TSeg orchestrates this and makes sure the arguments and parameters are passed to the corresponding selected segmentation tool properly and the execution will be handled accordingly. The parameters include but are not limited to input data location, output directory, and desired segmentation algorithm - Figure 2.4. This allows the end-user complete control over the process and feedback from each step of the process. The preprocessed images and relevant parameters are sent to a modular segmentation controller script. As an effort to allow future development on TSeg, the segmentation controller script shows how the pipeline integrates two completely different segmentation packages.

Tracking

The tracking widget of TSeg employs connected component analysis and the Hungarian algorithm for accurate cell tracking across 3D time-lapse images, and, leverages autoregressive modeling to analyze cell trajectories, enabling these trajectories to be clustered in an unsupervised manner for a deeper understanding of motility - Figure 2.6. Features in each segmented image are found using the scipy label function. In order to reduce any leftover noise, any features under a minimum size are filtered out and considered leftover noise. After feature extraction, centroids are calculated using the center of mass function in scipy. The centroid of the 3D cell can be used as a representation of the entire body during tracking. The tracking algorithm goes through each captured time instance and connects centroids to the likely next movement of the cell. Tracking involves a series of measures in order to avoid incorrect assignments. An incorrect assignment could lead to inaccurate result sets and unrealistic motility patterns. If the same number of features in each frame of time could be guaranteed from segmentation, minimum distance could assign features rather accurately. Since this is not a guarantee, the Hungarian algorithm must be used to associate a cost with the assignment of feature tracking. The Hungarian method is a combinatorial optimization

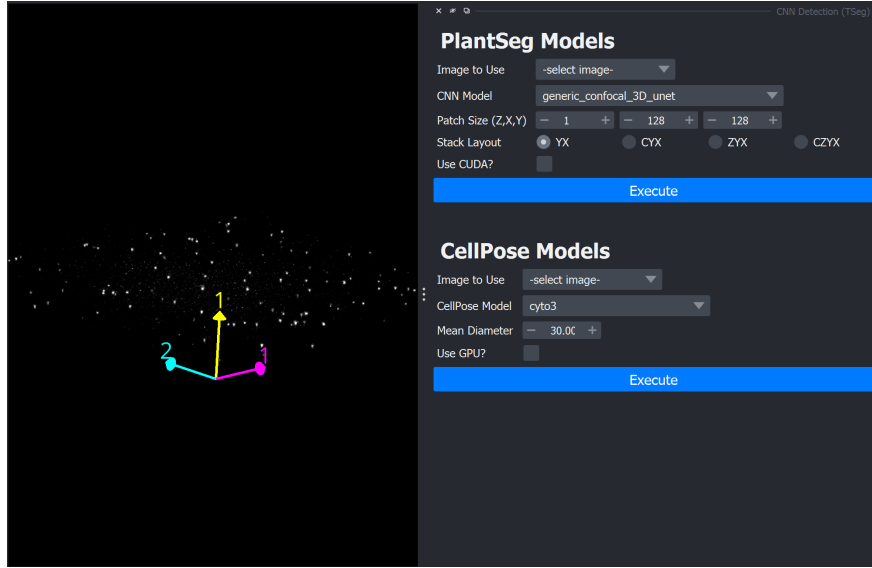


Figure 2.4: The CNN Detection widget integrates PlantSeg for tissue-specific 3D segmentation and CellPose for diverse cell types. These tools are implemented in the backend via their APIs, ensuring seamless operation.

algorithm that solves the assignment problem in polynomial time. The cost for the tracking algorithm determines which feature is the next iteration of the cell's tracking through the complete time series. The combination of distance between centroids for all previous points and the distance to the potential new centroid. If an optimal next centroid cannot be found within an acceptable distance of the current point, the tracking for the cell is considered as complete. Likewise, if a feature is not assigned to a current centroid, this feature is considered a new object and is tracked as the algorithm progresses. The complete path for each feature is then stored for motility analysis.

Motion Classification

To classify the motility pattern of *T. gondii* in 3D space in an unsupervised fashion we implement and use the method that Fazli et. al. introduced [21]. In that work, they used an autoregressive model (AR); a linear dynamical system that encodes a Markov-based transition prediction method. The reason is that although K-means is a favorable clustering algorithm, there are a few drawbacks to it and to the conventional methods that draw them impractical. Firstly, K-means assumes Euclidian distance, but AR motion parameters are geodesics that do not reside in a Euclidean space, and secondly, K-means assumes isotropic clusters, however, although AR motion parameters may exhibit isotropy in their space, without a proper distance metric, this issue cannot be clearly examined [21].

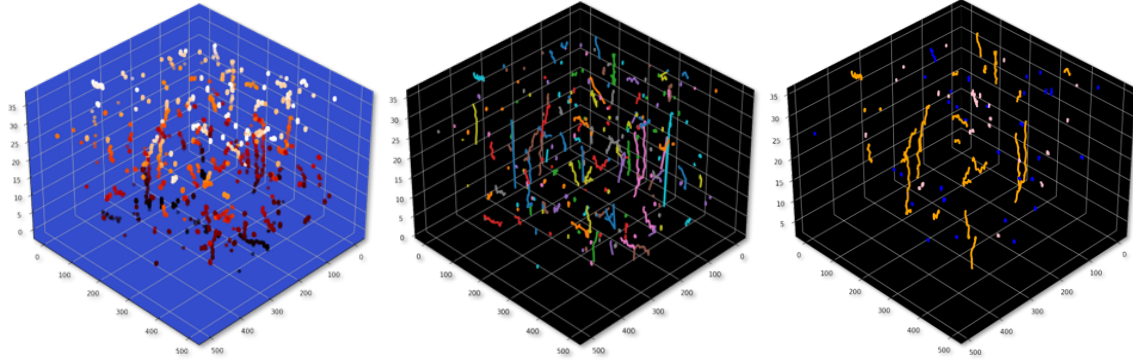


Figure 2.5: Left, 3D connected component labeling (CCL) is used to extract features from the segmented images. Middle, the centroids of the features are calculated using the center of mass function in scipy. Right, the tracking algorithm connects centroids across time instances to track the cells.

2.4 Evaluation

TSeg’s performance in segmentation was evaluated over datasets introduced in [68]. The Cell Tracking Challenge (CTC) offers a diverse array of 2D and 3D time-lapse microscopy datasets, each capturing unique biological specimens under various imaging modalities. Table A.1 contains an overview of these datasets, detailing the organisms studied, imaging techniques employed, and acquisition specifics. Table A.2 shows a sample image of each dataset. CellPose has 26 and PlantSeg has 17 different pre-trained models that can perform segmentation over 2D and 3D biomedical data. 10 samples from the 2D datasets, and one from the 3D datasets were randomly selected and processed with each of the 43 models using the API provided by PlantSeg and CellPose. Each dataset contains sequences of time-lapse video frames, therefore sample of a 2D dataset is comprised of a single 2D grayscale image, and each sample from the 3D datasets has a stack of 2D images recorded simultaneously across the z-axis to comprise one frame. The predicted masks of the models were evaluated against the provided ground-truth data using the Jaccard Index (JI) score and averaged across all samples of the same dataset. The results of the best performing models of CellPose and PlantSeg are shown in Table 2.2 and Table 2.3 over 2D and 3D datasets respectively.

The JI scores of all models are available in the appendix. Tables A.3 and A.4 present the performance metrics for all CellPose and PlantSeg models on the 2D datasets. Tables A.5 and A.6 summarize their performance on the 3D datasets.

CellPose demonstrated consistently strong performance across the majority of the evaluated 2D datasets, achieving accuracy levels often exceeding 0.90. Datasets like *BF-C2DL-HSC*, *BF-C2DL-MuSC*, and *PhC-C2DL-PSC* showed particularly high segmentation accuracy, consistently surpassing 0.95. Notably, CellPose also maintained robust accuracy on diverse cell types and imaging modalities, indicating its effective generalization across different contexts.

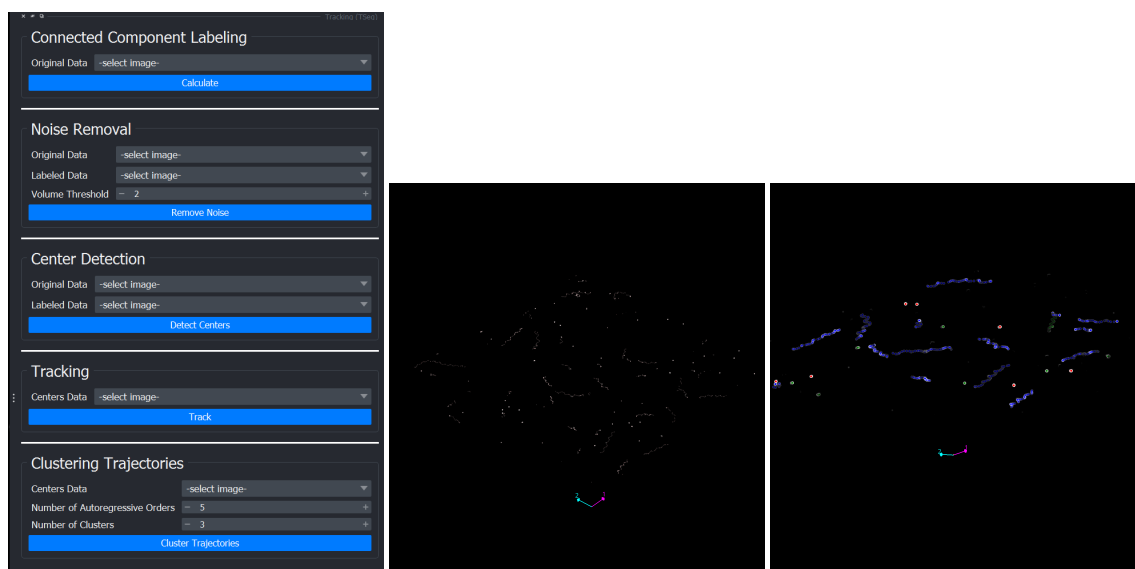


Figure 2.6: The tracking widget allows the user to set the parameters for the tracking and clustering algorithms and visualizes the results.

PlantSeg, evaluated primarily on plant-based imaging datasets, showed impressive robustness and consistently high accuracy, especially for structured cellular patterns. On both 2D and 3D datasets, PlantSeg achieved similar high-performance metrics to CellPose, often with accuracy scores around 0.96 or higher.

However, during evaluation on 3D datasets, PlantSeg encountered challenges due to computational constraints. Several large-sized samples could not be fully processed by specific CNN architectures, resulting in incomplete results (indicated by blank cells in Table A.6). This limitation primarily arose in datasets such as *Fluo-N₃DH-CE* and *Fluo-N₃DH-SIM+*, where the sheer volume of the input data surpassed the available computational resources, highlighting the computational demands associated with sophisticated 3D segmentation tasks. Future enhancements of PlantSeg may involve optimizing CNN architectures or leveraging computational strategies, such as tiling or cloud processing, to address these scalability limitations.

Overall, the evaluations demonstrate that both CellPose and PlantSeg effectively perform cell segmentation tasks across a range of datasets and conditions, with CellPose providing a slightly more versatile and robust generalization across various cell types and PlantSeg excelling particularly in structured datasets. Understanding these limitations and strengths will inform users' selection of the most appropriate tool based on their dataset characteristics and computational constraints. A comprehensive summary of segmentation performance metrics highlighting key comparative results between CellPose and PlantSeg on selected 2D and 3D datasets is presented in Tables 2.2 and 2.3. These tables succinctly capture essential findings, facilitating quick comparisons and reinforcing the suitability of TSeg's modular design in addressing diverse biomedical segmentation challenges. Furthermore, table 2.1 provides the Jaccard Index of the

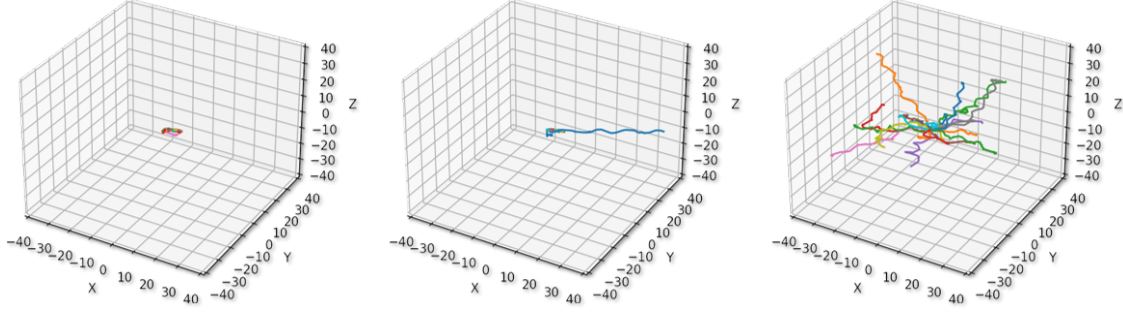


Figure 2.7: Clustering of *T. gondii* motility patterns in 3D space using an autoregressive model (AR) as introduced by Fazli et al. [21]. The AR model addresses the limitations of K-means by considering geodesic distances and non-isotropic clusters.

top performing models from the Cell Tracking Challenge (CTC) 2023¹, showcasing the state-of-the-art performance of TSeg in comparison to other leading segmentation tools.

2.5 Conclusion and Final Remarks

In this work, we presented TSeg, an intuitive and comprehensive pipeline specifically designed for the segmentation, tracking, and motility clustering of *Toxoplasma gondii* in 3D microscopic imaging data. Our pipeline leverages autoregressive parameterization to effectively capture temporal dependencies, successfully identifying distinct motility patterns and accurately clustering them based on their inherent characteristics. This approach provides insights into the complex motility behavior exhibited by *T. gondii* throughout its lytic cycle, facilitating more detailed and quantitative analyses compared to conventional methods.

One of the key contributions of our work is the seamless integration of state-of-the-art deep learning methodologies into an accessible and user-friendly platform. By incorporating powerful segmentation and detection tools such as CellPose and PlantSeg into TSeg, users benefit from high-performing deep learning algorithms without requiring extensive expertise in deep learning or computer vision frameworks. Additionally, our implementation of TSeg as a plugin for the Napari viewer further simplifies its adoption, providing interactive visualization and intuitive manipulation of 3D microscopy data.

TSeg significantly enhances the daily workflow of biological researchers by democratizing the analysis of complex 3D microscopic image data, making sophisticated quantitative studies accessible even to users without deep learning or extensive programming knowledge. Its implementation as a Napari plugin provides an intuitive graphical interface featuring immediate 3D visualization and feedback at each stage, allowing users to iteratively refine parameters and evaluate intermediate results. This user-centric

¹More information about the top-performing algorithms can be found <https://celltrackingchallenge.net/latest-csb-results/>

Table 2.1: Top-Performing Jaccard Index Scores on 2D CTC Datasets.

Dataset/Metric	Jaccard Index
BF-C2DL-HSC	0.855
BF-C2DL-MuSC	0.784
DIC-C2DH-HeLa	0.877
Fluo-C2DL-Huh7	0.811
Fluo-C2DL-MSC	0.687
Fluo-C3DH-A549	0.908
Fluo-C3DH-H157	0.890
Fluo-C3DL-MDA231	0.710
Fluo-N2DH-GOWT1	0.938
Fluo-N2DL-HeLa	0.923
Fluo-N3DH-CE	0.759
Fluo-N3DH-CHO	0.925
Fluo-N3DL-DRO	0.760
Fluo-N3DL-TRIC	0.821
Fluo-N3DL-TRIF	0.793
PhC-C2DH-U373	0.931
PhC-C2DL-PSC	0.756
Fluo-C3D-A549-SIM	0.955
Fluo-N2DH-SIM+	0.832
Fluo-N3DH-SIM+	0.906

More information about the top-performing algorithms can be found <https://celltrackingchallenge.net/latest-csb-results/>

design, combined with powerful underlying algorithms leveraging state-of-the-art tools like CellPose and PlantSeg, substantially improves the speed and quality of cell segmentation and tracking compared to manual methods or pipelines requiring extensive customization, directly addressing the challenges of time-consuming and expertise-heavy analysis.

Furthermore, TSeg functions as an adhesive component within the broader open-source scientific ecosystem. Built upon the familiar SciPy stack and integrated directly into Napari, its utility extends beyond its initial focus on *T. gondii*, proving adaptable to various other cell types and organisms. Its development exemplifies open-science principles through its open-source availability and reliance on well-regarded packages, aligning with the collaborative spirit fostered by communities such as PyOpenSci and the Journal of Open Source Software (JOSS). Looking forward, continuous enhancements in computational efficiency and scalability are envisioned to further broaden TSeg’s usability across larger and more complex datasets, ultimately fulfilling its significant potential to accelerate biological research by making advanced analyses readily available to a wider scientific audience.

Table 2.2: Best Segmentation Performance on 2D CTC Datasets (Jaccard Index)

Dataset Name	JI (Best CellPose Model(s))	JI (Best PlantSeg Model)	JI (CTC Benchmark)
BF-C2DL-HSC	0.99 (nuclei / yeast_BF_cp3)	0.99 (*)	0.855
BF-C2DL-MuSC	0.99 (nuclei / livecell_cp3)	0.99 (*)	0.784
DIC-C2DH-HeLa	0.36 (cyto3 / nuclei)	0.36 (*)	0.877
Fluo-C2DL-Huh7	0.60 (cyto3 / nuclei)	0.59 (*)	0.811
Fluo-C2DL-MSC	0.90 (cyto3 / cyto2)	0.89 (*)	0.687
Fluo-N2DH-GOWT1	0.89 (cyto3)	0.86 (*)	0.938
Fluo-N2DH-SIM+	0.89 (neurips_grayscale_cyto2)	0.80 (*)	0.832
Fluo-N2DL-HeLa	0.75 (cyto3 / nuclei)	0.75 (*)	0.923
PhC-C2DH-U373	0.87 (cyto3 / nuclei)	0.87 (*)	0.931
PhC-C2DL-PSC	0.91 (tissuenet_cp3 / livecell_cp3)	0.91 (*)	0.756

(*) Indicates confocal_2D_unet_ovules_ds2x.

Table 2.3: Best Segmentation Performance on 3D CTC Datasets (Jaccard Index)

Dataset Name	JI (Best CellPose Models)	JI (Best PlantSeg Model)	JI (CTC Benchmark)
Fluo-C3DH-A549	0.98 (CPx, CP)	0.96 (*)	0.908
Fluo-C3DH-A549-SIM	0.99 (CPx, CP, TN3)	0.97 (*)	0.955
Fluo-C3DH-H157	0.93 (cyto3, nuclei, cyto2_cp3)	0.88 (*)	0.890
Fluo-N3DH-CE	0.80 (cyto3, nuclei, cyto2_cp3)	0.78 (*)	0.759
Fluo-N3DH-CHO	0.84 (nuclei, tissuenet_cp3)	0.84 (*)	0.925
Fluo-N3DH-SIM+	0.94 (cyto2)	— (N/A (Processing failed))	0.906

(*) Indicates generic_confocal_3D_unet

2.6 Limitations and Future Directions

While TSeg provides an intuitive and integrated pipeline for 3D cell analysis, several limitations should be acknowledged.

Limitations of TSeg:

- Its performance can face computational constraints when processing large 3D datasets. This is particularly true when utilizing integrated tools like PlantSeg, which sometimes encountered memory limitations on complex CTC samples. Consequently, TSeg’s overall accuracy is inherently linked to the performance and limitations of the specific tools it incorporates, namely PlantSeg and CellPose.

- The tracking module, relying on centroid calculation and the Hungarian algorithm, may encounter difficulties with complex cellular events such as division, fusion, or temporary occlusion. This could potentially lead to assignment errors in dense or highly dynamic scenarios.
- Although designed with generic functions and evaluated on various cell types, TSeg was initially developed focusing on *T. gondii*. Its effectiveness might therefore vary on significantly different cell types or imaging modalities not covered in the evaluations, requiring further validation for broader generalization.
- Furthermore, the motility classification employs an AR model which might not capture the full complexity of diverse biological movement patterns observed in different cell types or conditions.
- Finally, while the preprocessing module addresses common artifacts like vignetting and noise, it may not encompass all specific artifacts encountered in varied microscopy setups, potentially necessitating external preprocessing steps for optimal results.

Future Directions:

- Addressing the computational bottlenecks to improve scalability and efficiency, especially for large 3D time-lapse datasets, is a key priority for future development. This could involve optimizing existing algorithms or exploring strategies like data tiling.
- Expanding the repertoire of integrated segmentation and tracking algorithms beyond CellPose and PlantSeg can offer users greater flexibility. Incorporating more diverse algorithms could provide potentially improved performance tailored to specific biological contexts or challenging imaging conditions.
- Enhancing the robustness of the tracking algorithm is crucial for broader applicability. Future work should focus on better handling complex events like high cell density, cell division, fusion events, and temporary disappearance or occlusion of cells, perhaps exploring alternative tracking paradigms. Improving the tracking module could enable TSeg to handle more complex scenarios, such as mitochondrial dynamics or cell-cell interactions, which are common in biological systems.
- Advancing the motion classification module by exploring non-linear models or alternative machine learning approaches could provide deeper insights into complex motility phenotypes beyond the capabilities of the current AR model. Furthermore, exploring the spectrum of motion of parameterized trajectories could yield valuable insights into the underlying biological processes if a true correspondance between the motility phenotypes and the motion manifold is established.
- More extensive validation of TSeg's generalizability across a wider range of 3D cell types, organisms, and imaging modalities is needed to better define its scope and reliability in diverse research settings.
- Integrating TSeg's outputs seamlessly with downstream quantitative analysis tools would facilitate more comprehensive biological investigations, allowing researchers to easily move from segmentation and tracking to deeper statistical analysis.

- Lastly, incorporating mechanisms for user-guided refinement within the Napari interface could improve usability. Allowing users to interactively correct segmentation or tracking results would be particularly beneficial in ambiguous or challenging cases.

CHAPTER 3

TRAINING A SUPERVISED CILIA SEGMENTATION MODEL FROM SELF-SUPERVISION

3.1 Introduction

Cilia are hair-like membranes that extend out from the surface of the cells and are present on a variety of cell types such as lungs and brain ventricles and can be found in the majority of vertebrate cells. Categorized into motile and primary, motile cilia can help the cell to propel, move the flow of fluid, or fulfill sensory functions, while primary cilia act as signal receivers, translating extracellular signals into cellular responses [30]. Ciliopathies is the term commonly used to describe diseases caused by ciliary dysfunction. These disorders can result in serious issues such as blindness, neurodevelopmental defects, or obesity [25]. Motile cilia beat in a coordinated manner with a specific frequency and pattern [56]. Stationary, dyskinetic, or slow ciliary beating indicates ciliary defects. Ciliary beating is a fundamental biological process that is essential for the proper functioning of various organs, which makes understanding the ciliary phenotypes a crucial step towards understanding ciliopathies and the conditions stemming from it [109].

Identifying and categorizing the motion of cilia is an essential step towards understanding ciliopathies. However, this is generally an expert-intensive process. Studies have proposed methods that automate the ciliary motion assessment [110]. These methods rely on large amounts of labeled data that are annotated manually which is a costly, time-consuming, and error-prone task. Consequently, a significant bottleneck to automating cilia analysis is a lack of automated segmentation. Segmentation has remained a bottleneck of the pipeline due to the poor performance of even state-of-the-art models on some datasets. These datasets tend to exhibit significant spatial artifacts (light diffraction, out-of-focus cells, etc.) which confuse traditional image segmentation models [63].

Video segmentation techniques tend to be more robust to such noise, but still struggle due to the wild inconsistencies in cilia behavior: while healthy cilia have regular and predictable movements, unhealthy cilia display a wide range of motion, including a lack of motion altogether [43]. This lack of motion

especially confounds movement-based methods which otherwise have no way of discerning the cilia from other non-cilia parts of the video. Both image and video segmentation techniques tend to require expert-labeled ground truth segmentation masks. Image segmentation requires the masks in order to effectively train neural segmentation models to recognize cilia, rather than other spurious textures. Video segmentation, by contrast, requires these masks in order to properly recognize both healthy and diseased cilia as a single cilia category, especially when the cilia show no movement.

To address this challenge, we propose a two-stage image segmentation model designed to obviate the need for expert-drawn masks. We first build a corpus of segmentation masks based on optical flow (OF) thresholding over a subset of healthy training data with guaranteed motility. We then train a semi-supervised neural segmentation model to identify both motile and immotile data as a single segmentation category, using the flow-generated masks as “pseudo-labels”. These pseudo-labels operate as “ground truth” for the model while acknowledging the intrinsic uncertainty of the labels. The fact that motile and immotile cilia tend to be visually similar in snapshot allows us to generalize the domain of the model from motile cilia to all cilia. Combining these stages results in a semi-supervised framework that does not rely on any expert-drawn ground-truth segmentation masks, paving the way for full automation of a general cilia analysis pipeline.

3.2 Background

Dysfunction in ciliary motion indicates diseases known as ciliopathies, which can disrupt the functionality of critical organs like the lungs and kidneys. Understanding ciliary motion is crucial for diagnosing and understanding these conditions. The development of diagnosis and treatment requires the measurement of different cell properties including size, shape, and motility [92].

Accurate analysis of ciliary motion is essential but challenging due to the limitations of manual analysis, which is labor-intensive, subjective, and prone to error. [110] proposed a modular generative pipeline that automates ciliary motion analysis by segmenting, representing, and modeling the dynamic behavior of cilia, thereby reducing the need for expert intervention and improving diagnostic consistency. [75] developed a computational pipeline using dynamic texture analysis and machine learning to objectively and quantitatively assess ciliary motion, achieving over 90% classification accuracy in identifying abnormal ciliary motion associated with diseases like primary ciliary dyskinesia (PCD). Additionally, [109] explored advanced feature extraction techniques like Zero-phase PCA Sphering (ZCA) and Sparse Autoencoders (SAE) to enhance cilia segmentation accuracy. These methods address challenges posed by noisy, partially occluded, and out-of-phase imagery, ultimately improving the overall performance of ciliary motion analysis pipelines. Collectively, these approaches aim to enhance diagnostic accuracy and efficiency, making ciliary motion analysis more accessible and reliable, thereby improving patient outcomes through early and accurate detection of ciliopathies. However, these studies rely on manually labeled data. The segmentation masks and ground-truth annotations, which are essential for training the models and validating their performance, are generated by expert reviewers. This dependence on manually labeled data is a significant limitation making automated cilia segmentation the bottleneck to automating cilia analysis.

In the biomedical field, where labeled data is often scarce and costly to obtain, several solutions have been proposed to augment and utilize available data effectively. These include semi-supervised learning [95], [104], which utilizes both labeled and unlabeled data to enhance learning accuracy by leveraging the data’s underlying distribution. Active learning [86] focuses on selectively querying the most informative data points for expert labeling, optimizing the training process by using the most valuable examples. Data augmentation techniques [8], [53], [54], [81], [84], [94], [104], [105], such as image transformations and synthetic data generation through Generative Adversarial Networks [22], [108], increase the diversity and volume of training data, enhancing model robustness and reducing overfitting. Transfer learning [35], [77], [85], [104] transfers knowledge from one task to another, minimizing the need for extensive labeled data in new tasks. Self-supervised learning [45], [52], [66] creates its labels by defining a pretext task, like predicting the position of a randomly cropped image patch, aiding in the learning of useful data representations. Additionally, few-shot, one-shot, and zero-shot learning techniques [58], [70] are designed to operate with minimal or no labeled examples, relying on generalization capabilities or metadata for making predictions about unseen classes.

A promising approach to overcome the dependency on manually labeled data is the use of unsupervised methods to generate ground truth masks. Unsupervised methods do not require prior knowledge of the data [44]. Using domain-specific cues unsupervised learning techniques can automatically discover patterns and structures in the data without the need for labeled examples, potentially simplifying the process of generating accurate segmentation masks for cilia. Inspired by advances in unsupervised methods for image segmentation, in this work, we firstly compute the motion vectors using optical flow of the ciliary regions and then apply autoregressive modelling to capture their temporal dynamics. Autoregressive modelling is advantageous since the labels are features themselves. By analyzing the OF vectors, we can identify the characteristic motion of cilia, which allows us to generate pseudo-labels as ground truth segmentation masks. These pseudo-labels are then used to train a robust semi-supervised neural network, enabling accurate and automated segmentation of both motile and immotile cilia.

3.3 Methodology

Dynamic textures, such as sea waves, smoke, and foliage, are sequences of images of moving scenes that exhibit certain stationarity properties in time [16]. Similarly, ciliary motion can be considered as dynamic textures for their orderly rhythmic beating. Taking advantage of this temporal regularity in ciliary motion, OF can be used to compute the flow vectors of each pixel of high-speed videos of cilia. In conjunction with OF, autoregressive (AR) parameterization of the OF property of the video yields a manifold that quantifies the characteristic motion in the cilia. The low dimension of this manifold contains the majority of variations within the data, which can then be used to segment the motile ciliary regions.

3.3.1 Optical Flow Properties

Taking advantage of this temporal regularity in ciliary motion, we use OF to capture the motion vectors of ciliary regions in high-speed videos. OF provides the horizontal u and vertical v components of the motion for each pixel. From these motion vectors, several components can be derived such as the magnitude, direction, divergence, and importantly, the curl (rotation). The curl, in this context, represents the rotational motion of the cilia, which is indicative of their rhythmic beating patterns. We extract flow vectors of the video recording of cilia, under the assumption that pixel intensity remains constant throughout the video.

$$I(x, y, t) = I(x + u\delta t, y + v\delta t, t + \delta t) \quad (3.3.1)$$

(3.3.1) Where $I_{x,y,t}$ is the pixel intensity at position x, y a time t . Here, u_t, v_t are small changes in the next frame taken after t time, and u, v , respectively, are the OF components that represent the displacement in pixel positions between consecutive frames in the horizontal and vertical directions at pixel location x, y .

3.3.2 Autoregressive Modeling

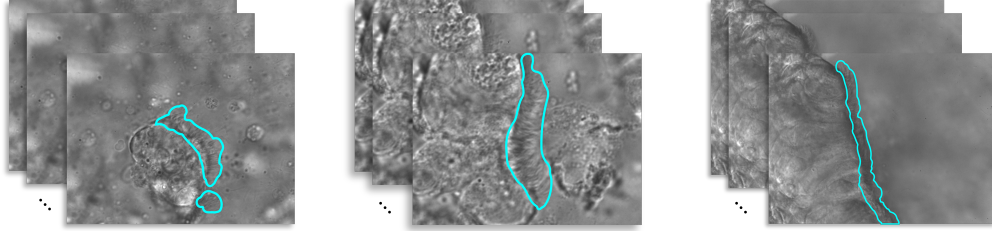


Figure 3.1: A sample of three videos in our cilia dataset with their manually annotated ground truth masks.

Figure 3.1 shows a sample of the OF component at a random time. From OF vectors, elemental components such as rotation are derived, which highlights the ciliary motion by capturing twisting and turning movements. To model the temporal evolution of these motion vectors, we employ an autoregressive (AR) model [36]. This model captures the dynamics of the flow vectors over time, allowing us to understand how the motion evolves frame by frame. The AR model helps in decomposing the motion into a low-dimensional subspace, which simplifies the complex ciliary motion into more manageable analyses.

$$y_t = C\vec{x}_t + \vec{u} \quad (3.3.2)$$

$$\vec{x}_t = A_1\vec{x}_{t-1} + A_2\vec{x}_{t-2} + \dots + A_d\vec{x}_{t-d} + \vec{v}_t \quad (3.3.3)$$

In equation (3.3.2), y_t represents the appearance of cilia at time t influenced by noise u . Equation (3.3.3) represents the state x of the ciliary motion in a low-dimensional subspace defined by an orthogonal basis C at time t , plus a noise term v_t and how the state changes from t to $t + 1$.

Equation (3.3.3) is a decomposition of each frame of a ciliary motion video y_t into a low-dimensional state vector x_t using an orthogonal basis C . This equation at position x_t is a function of the sum of d of its previous positions $x_{t-1}, x_{t-2}, x_{t-d}$ each multiplied by its corresponding coefficients $A = A_1, A_2, \dots, A_d$. The noise terms u and v are used to represent the residual difference between the observed data and the solutions to the linear equations. The variance in the data is predominantly captured by a few dimensions of C , simplifying the complex motion into manageable analyses.

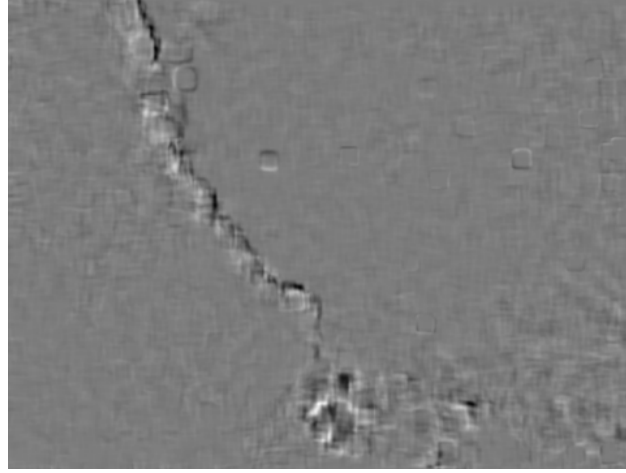


Figure 3.2: Representation of rotation (curl) component of OF at a random time.

Each order of the autoregressive model roughly aligns with different frequencies within the data, therefore, in our experiments, we chose $d=5$ as the order of our autoregressive model. This choice allows us to capture a broader temporal context, providing a more comprehensive understanding of the system's dynamics. We then created raw masks from this lower-dimensional subspace, and further enhanced them with adaptive thresholding to remove the remaining noise.

In 3.2, the first-order AR parameter is showing the most variance in the video, which corresponds to the frequency of motion that cilia exhibit. The remaining orders have correspondence with other different frequencies in the data caused by, for instance, camera shaking. Evidently, simply thresholding the first-order AR parameter is adequate to produce an accurate mask, however, in order to get a more refined result we subtracted the second order from the first one, followed by a Min-Max normalization of pixel intensities and scaling to an 8-bit unsigned integer range. We used adaptive thresholding to extract the mask on all videos of our dataset. The generated masks exhibited under-segmentation in the ciliary region, and sparse over-segmentation in other regions of the image. To overcome this, we adapted a Gaussian blur filter followed by an Otsu thresholding to restore the under-segmentation and remove the sparse over-segmentation. Figure 3.4 illustrates the steps of the process.

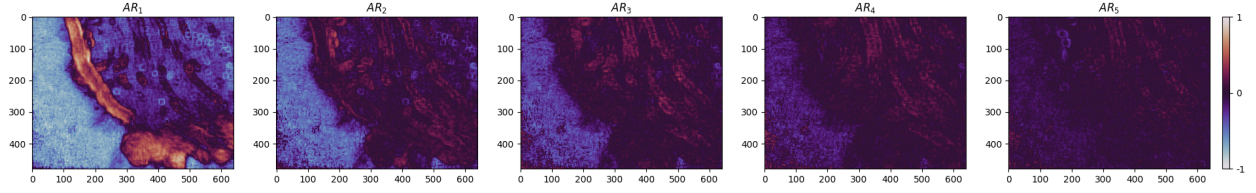


Figure 3.3: The pixel representation of the 5-order AR model of the OF component of a sample video. The x and y axes correspond to the width and height of the video.

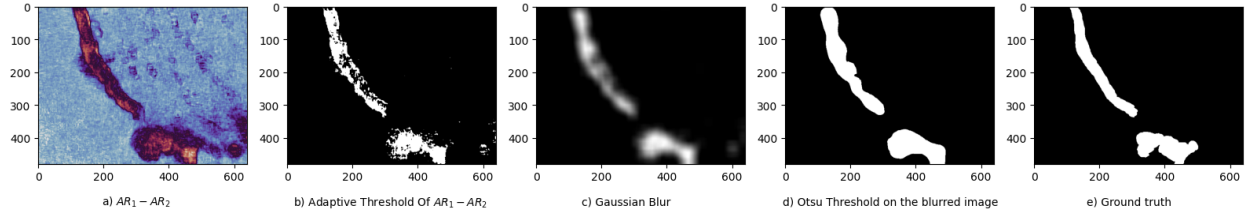


Figure 3.4: The process of computing the masks. **a)** Subtracting the second-order AR parameter from the first-order, followed by **b)** Adaptive thresholding, which suffers from under/over-segmentation. **c)** A Gaussian blur filter, followed by **d)** An Otsu thresholding eliminates the under/over-segmentation.

3.3.3 Training the model

Our dataset includes 512 videos, with 437 videos of dyskinetic cilia and 75 videos of healthy motile cilia, referred to as the control group. The control group is split into %85 and %15 for training and validation respectively. 108 videos in the dyskinetic group are manually annotated which are used in the testing step. Figure 3.1 shows annotated samples of our dataset.

In our study, we employed a Feature Pyramid Network (FPN) [48] architecture with a ResNet-34 encoder. The model was configured to handle grayscale images with a single input channel and produce binary segmentation masks. For the training input, one mask is generated per video using our methodology, and we use the first 250 frames from each video in the control group making a total of 18,750 input images. We utilized Binary Cross-Entropy Loss for training and the Adam optimizer with a learning rate of 10^{-3} . To evaluate the model's performance, we calculated the Dice score during training and validation. Data augmentation techniques, including resizing, random cropping, and rotation, were applied to enhance the model's generalization capability. The implementation was done using a library [37] based on PyTorch Lightning to facilitate efficient training and evaluation. Table 3.1 contains a summary of the model parameters and specifications.

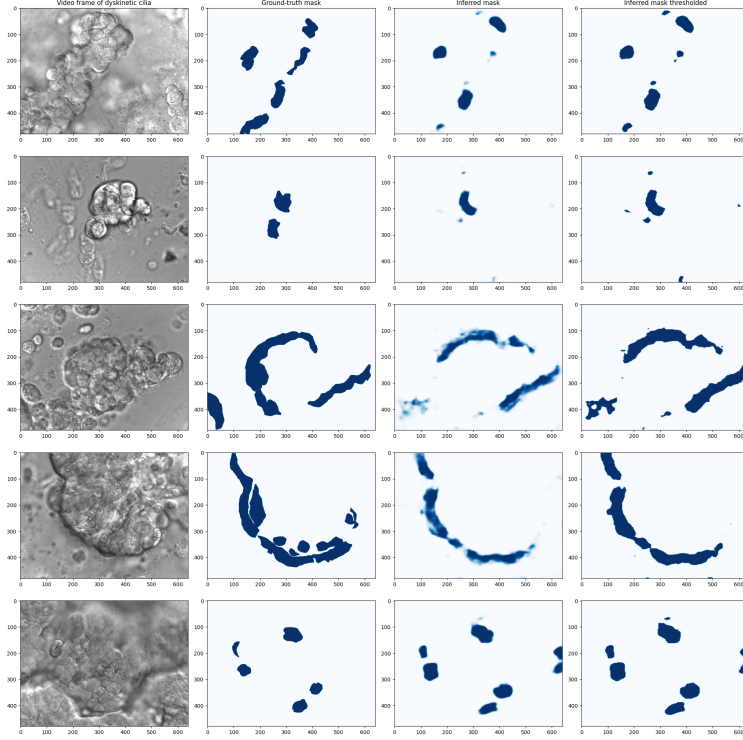


Figure 3.5: The model predictions on 5 dyskinetic cilia samples. The first column shows a frame of the video, the second column shows the manually labeled ground truth, the third column is the model’s prediction, and the last column is a thresholded version of the prediction.

3.4 Results and Discussion

The model’s performance metrics, including IoU, Dice score, sensitivity, and specificity, are summarized in Table 3.2. The validation phase achieved an IoU of 0.312 and a Dice score of 0.476, which indicates a moderate overlap between the predicted and ground truth masks. The high sensitivity (0.999) observed during validation suggests that the model is proficient in identifying ciliary regions, albeit with a specificity of 0.813, indicating some degree of false positives. In the testing phase, the IoU and Dice scores decreased to 0.230 and 0.374, respectively, reflecting the challenges posed by the dyskinetic cilia data, which were not included in the training or validation sets. Despite this, the model maintained a reasonable sensitivity of 0.631 and specificity of 0.787.

Figure 3.5 provides visual examples of the model’s predictions on dyskinetic cilia samples, alongside the manually labeled ground truth and thresholded predictions. The dyskinetic samples were not used in the training or validation phases. These predictions were generated after only 20 epochs of training with a small training data. The visual comparison reveals that, while the model captures the general structure of ciliary regions, there are instances of under-segmentation and over-segmentation, which are

Table 3.1: Summary of model architecture, training setup, and dataset distribution

Aspect	Details
Architecture	FPN with ResNet-34 encoder
Input	Grayscale video frame with a single input channel
Number of Epochs	20
Batch Size	4
Training Samples	64 videos
Validation Samples	11 videos
Test Samples	108 videos
Loss Function	Binary Cross-Entropy Loss
Optimizer	Adam optimizer with a learning rate of 10^{-3}
Evaluation Metric	Dice score during training and validation
Data Augmentation Techniques	Resizing, random cropping, and rotation
Implementation	Using a Python library with Neural Networks for Image Segmentation based on PyTorch [37]

more pronounced in the dyskinetic samples. This observation is consistent with the quantitative metrics, suggesting that further refinement of the pseudo-label generation process or model architecture could enhance segmentation accuracy.

The observed difference in performance metrics between the validation and testing phases (Table 3.2) is primarily due to the distinct nature of the datasets used. The model was trained and validated using pseudo-labels generated exclusively from the 75 motile (healthy) cilia videos. Conversely, the testing phase utilized a separate set of 108 dyskinetic cilia videos, presenting the model with a larger set of unseen samples. Another potential reason for the performance discrepancy lies in the differing visual presentation of cilia between the training/validation (motile) and testing (dyskinetic) datasets. In many videos of the healthy motile set, cilia might prominently extend from the cell surfaces, creating clear, high-contrast silhouettes against the background or adjacent cell borders. Consequently, the model may have learned a strong, potentially spurious, correlation associating these distinct border features with the presence of cilia. However, if the dyskinetic cilia in the test set are less clearly defined, visually blend more with surrounding tissue, or if similar border/edge features exist in the images without corresponding cilia, this learned association would be misleading. This reliance on potentially biased visual cues from the training data, rather than solely on the intrinsic appearance of cilia themselves, would likely contribute to segmentation errors (e.g., false positives at borders or false negatives for less distinct cilia) and the reduced performance observed during testing.

Table 3.2: The performance of the model in validation and testing phases.

Phases	IoU over dataset	Dice Score	Sensitivity	Specificity
Validation	0.312	0.476	0.999	0.813
Testing	0.230	0.374	0.631	0.787

These results show the potential of our approach to reduce the reliance on manually labeled data for cilia segmentation. The use of this unsupervised learning framework allows the model to generalize from the motile cilia domain to the more variable dyskinetic cilia, although with some limitations in accuracy. Future work could focus on expanding the dataset and improving the process of generating pseudo-labels to enhance the model’s accuracy.

Table 3.3: Comparison of Cilia Segmentation Performance Metrics with Zain et al. [109].

Metric	Validation	Testing	Zain et al. [109]
IoU (Jaccard Index)	0.312	0.230	0.441
Dice Score	0.476	0.374	0.585
Sensitivity (Recall)	0.999	0.631	0.624
Specificity	0.813	0.787	N/A

*Best result reported for the corresponding metric in Figure 10 of Zain et al.. Note that the best performance for each metric might come from slightly different model configurations within that study.

A comparison with the state-of-the-art results reported by Zain et al. [109] highlights the performance characteristics of our self-supervised approach (Table 3.3). While the overall segmentation overlap metrics achieved by our model during testing (IoU 0.230, Dice 0.374) are lower than the best scores obtained by the feature-enhanced supervised method of Zain et al. (IoU 0.441, Dice 0.585), our model demonstrates comparable performance in identifying true ciliary regions. Specifically, the sensitivity achieved on our dyskinetic test set (0.631) is notably similar to the best recall reported by Zain et al. (0.624). This suggests that while challenges remain in achieving precise spatial accuracy using self-supervision, particularly when generalizing from motile pseudolabels to dyskinetic cilia, our method effectively learns to recognize cilia presence at a rate comparable to enhanced supervised techniques.

3.5 Conclusions and Final Remarks

In this chapter, we introduced a self-supervised framework for cilia segmentation that obviates the need for expert-labeled ground truth masks. By leveraging the rhythmic motion signatures of motile cilia, we generated pseudo-labels from optical flow (OF) properties and an autoregressive model of the flow vectors. These pseudo-labels were then used to train a semi-supervised neural network on motile cilia, enabling the model to generalize to dyskinetic cilia without requiring additional annotations.

Key Contributions. A central contribution of this work lies in showing that motile and dyskinetic cilia, despite significantly different motion patterns, share enough visual similarity in static frames to be captured under a single segmentation category. Our motion-based pseudo-label creation removes a major bottleneck in cilia analysis namely, the laborious process of manually annotating ciliary structures in high-speed videos. By emphasizing a two-stage pipeline (unsupervised pseudo-label generation followed by semi-supervised training), we have reduced reliance on large labeled datasets and demonstrated a pathway for robust segmentation of both healthy and diseased cilia.

Discussion of Performance. Quantitative evaluations on dyskinetic cilia samples which the model never saw during training revealed moderate segmentation performance (IoU of 0.230 and Dice score of 0.374). While these results are lower than on the validation set (which contained only motile cilia), they underscore the model’s capacity to transfer knowledge from one domain to another. Visual inspection of predictions suggests that refining the pseudo-label creation process could further reduce false positives and improve segmentation boundaries, especially for dyskinetic cilia with complex or minimal motion signatures.

Limitations. Despite the progress achieved, several limitations persist. First, the OF-based pseudo-labels can be imperfect whenever extraneous motion (e.g., camera shake) mimics ciliary motion or when cilia exhibit barely detectable movement. Additionally, the AR modeling approach, although effective for regular rhythmic patterns, might struggle with extremely irregular or sporadic ciliary motion in advanced ciliopathies. Finally, our training dataset was comparatively small, which may limit model generalizability to more diverse clinical scenarios or imaging conditions.

Future Directions. Moving forward, the framework could be refined in multiple ways:

- **Improved pseudo-labels:** Incorporating additional video processing techniques (e.g., temporal filtering or background subtraction) could yield higher-quality pseudo-labels, thus reducing noise and boosting performance on dyskinetic cilia.
- **Data Augmentation and Expansion:** Given the variability in ciliary motion across different tissue types and disease states, assembling a larger and more diverse dataset including videos from multiple clinical centers would likely enhance the model’s robustness.
- **Advanced Architectures:** Exploring more sophisticated neural network architectures (e.g., transformers or advanced spatio-temporal models) might better capture the nuanced motion of dyskinetic cilia.
- **Clinical Integration:** Integrating the pipeline with downstream analysis tools for instance, automated frequency or beat-pattern quantification could streamline clinical workflows and facilitate early detection of ciliopathies.

Broader Implications. Beyond cilia analysis, this chapter illustrates how self-supervision can address similar bottlenecks in biomedical image analysis, where labeled data are scarce. By turning domain-specific cues (i.e., ciliary motion) into an automated labeling mechanism, we move closer to high-throughput analysis pipelines that can adapt to various imaging contexts. As the subsequent chapters will explore complementary approaches to segmentation and classification tasks, the concepts introduced here especially the coupling of domain knowledge with unsupervised strategies offer a blueprint for tackling the annotation scarcity that frequently hinders medical image processing.

In sum, the work presented in this chapter establishes a promising avenue for more efficient and accurate cilia segmentation. Our self-supervised approach demonstrates that motion and visual cues, when harnessed effectively, can reduce the need for manual annotation and bolster the performance of segmentation models in both research and clinical settings. While challenges remain, particularly in handling more complex dyskinetic patterns, the framework set forth here provides a foundation upon which future studies can refine and expand, ultimately advancing the automated analysis of ciliopathies and related conditions.

CHAPTER 4

MINIMALLY-SUPERVISED BIOMEDICAL IMAGE SEGMENTATION VIA CONTRASTIVE LEARNING

4.1 Introduction

Image segmentation is a fundamental process in many computer vision applications and is used to partition the image into separate regions. It is an essential part in various biomedical applications, including lesion and tumor detection and analysis, organ localization and identification, diagnosis and monitoring, and cell and tissue analysis. Similarly, image segmentation is a cornerstone of quantitative cell research, particularly for studying cellular dynamics like motility [92] and morphological changes. Given its critical role, it has been the focus of extensive research, with ongoing advancements aimed at improving accuracy, automation, and generalization across diverse imaging modalities.

Biomedical images come in a vast variety of formats, types, and modalities [62], [96], [115]. Similarly, due to the variety of biological structures, segmentation targets can vary from nuclei and cell membranes to organelles such as mitochondria, cilia, tumors, and lesions, as well as blood vessels, bone, and brain structures [93]. Deep learning (DL) has advanced the field of image segmentation, particularly with the success of convolutional neural networks (CNN) [100]. While CNNs revolutionized segmentation for their high accuracy, due to the large diversity in biomedical image modalities, formats, and structures as well as the scarcity of ground truth data, CNNs are tailored for specific tasks [64] in biomedical image segmentation and therefore suffer from overfitting and exhibit poor generalizability over unseen data. Furthermore, their specificity to tasks, high computational demands, and complex implementation limit their broader application.

Inspired by Large Language Models (LLMs), Foundation Models such as the Segment Anything Model (SAM) [49] demonstrate excellent zero-shot segmentation performance across a large variety of general images. Studies that build upon SAM [46] have shown promising zero-shot learning capabilities and can segment objects in biomedical images regardless of their modality. However, when applied

to biomedical data without fine-tuning, SAM often struggles to match the accuracy of domain-specific models like U-Net. Its zero-shot performance varies significantly across medical datasets and tasks, highlighting the need for fine-tuning to adapt it effectively for biomedical image segmentation. Furthermore, although SAM excels at segmenting objects with well-defined, envelope or convex geometries, it struggles with biological structures that exhibit diffuse or punctate patterns such as cilia, which are even difficult to generate hand-drawn labels for.

Unsupervised methods, on the other hand, are used in scenarios where domain-specific cues suffice for crafting an algorithm for segmentation and when obtaining ground truth data is costly [93]. However, since they are domain-specific, unsupervised methods also exhibit poor generalizability. Self-supervised learning is also a promising direction in unsupervised segmentation. Contrastive learning (CL) is a successful variant of Self-supervised learning and refers to a type of learning where the goal is to learn representations by contrasting positive pairs (similar or related data points) against negative pairs (dissimilar or unrelated data points). This approach is widely used in self-supervised learning where labels are not available. Contrastive coding (CC), often seen as a subset or a specific implementation of CL, refers more specifically to the encoding process where contrastive loss functions are used to train models to produce these discriminative embeddings.

Contrastive learning provides an alternative approach to segmentation by leveraging similarities and differences in the data rather than relying on explicit labels. For addressing all the aforementioned issues, we turned to contrastive coding to teach the network to recognize objects of the same texture and configuration. By learning representations that cluster visually similar structures together while separating dissimilar ones, contrastive learning enables segmentation with minimal user interaction. This makes it particularly suitable for biomedical image analysis, where labeled data is scarce, and manual annotations are costly and time-consuming. The code to our method is available at <https://github.com/quinnngroup/contrastive-coding>

4.2 Background

Image segmentation is a crucial topic in computer vision and in particular deep learning. In image segmentation, an input image is broken down into its mutually exclusive semantic constituents such as the independent objects and the background. To address the first aforementioned shortcoming, Hyunseob et al. proposed a model called MDNet [72]. MDNet, or Multi Domain Network, is a supervised tracking method that learns domain-independent representations from pre-training. In supervised learning, a set including different objects which are semantically similar together, such as "pedestrian", "ball", "car", or "flower," are used for training. The most important drawback of MDNet, and any supervised segmentation framework like Mask- and Cascade-RCNN [5], [29] SSD [60], is that it relies on large amounts of labeled data for training from various objects, while after training the model with such a big dataset, there is still no guarantee that it can detect any other objects. There always exists sets of objects which are not used for training, and as a consequence, the network may not detect those types of objects properly.

Some unsupervised segmentation methods were proposed recently [15], [40], [61], [103], the most popular of which is W-Net [103]. However, since W-Net has to reconstruct the entire image again from the segmentation map, background and other objects of no interest have to be present in the segmentation mask, increasing the burden on the network to perfectly segment them, while they are of no interest. Indeed, we first started our experimentation to try to extend the work in W-Net, but we found it relied heavily on the final Conditional Random Field (CRF) module to fix the background creeping, and we could not distill the objects of interest alone with the segmentation mask. Also, in [15], an unsupervised segmentation method was proposed for separating the background from foreground using deep learning, and again, distilling the object of interest alone is still an issue in this work, since sometimes the object of interest is visually closer to the background than to the foreground, as in some of our data that we present later.

For addressing all the aforementioned issues we turned to contrastive coding in order to teach the network to recognize objects of the same texture and con guration. In contrastive coding, the goal is to represent instances (images/videos/patches) with vectors, and have instances that are similar attract and instances that are dissimilar repel each other. This is typically done with dot product or cosine similarity on the learnt vectors. There has been a lot of recent work in unsuper vised contrastive learning [12], [28], [74], such as SimCLR[10], [11], where patches from the same image are made to attract each other, while patches from different images are made to repel each other. To aid with better object recognition, the patches are transformed with the usual image augmentation techniques like color jittering, blurring, flipping, and rotation.

4.3 Methodology

4.3.1 Network architecture

The network is designed to take in a patch of dimensions $i \times k \times k$ and output a vector of size d to represent this patch in the dot product operations. We achieve this with a network constructed as follows: Three MBConv layers each outputting 32 channels, followed by a max pooling layer that downsizes the image by half, then another 3 MBConv layers each outputting 64 channels, followed by a global max pooling layer downsizing the image to $d \times 1 \times 1$ followed by a fully connected layer that outputs another d -dimensional vector with a final activation function of \tanh . Figure 4.1(a) illustrates the network architecture and the application of the network to the current and next video frames. An MBConv layer is adapted from EfficientNet[90]. Figure 4.1(b) shows the internals of an MBConv layer.

4.3.2 Contrastive training

Let x_1, x_2, \dots, x_N be patches of size $i \times k \times k$ pixels from an input image I_t . We aim to represent each patch with a vector representation of size d . The vector representation of a patch x_i is obtained using a

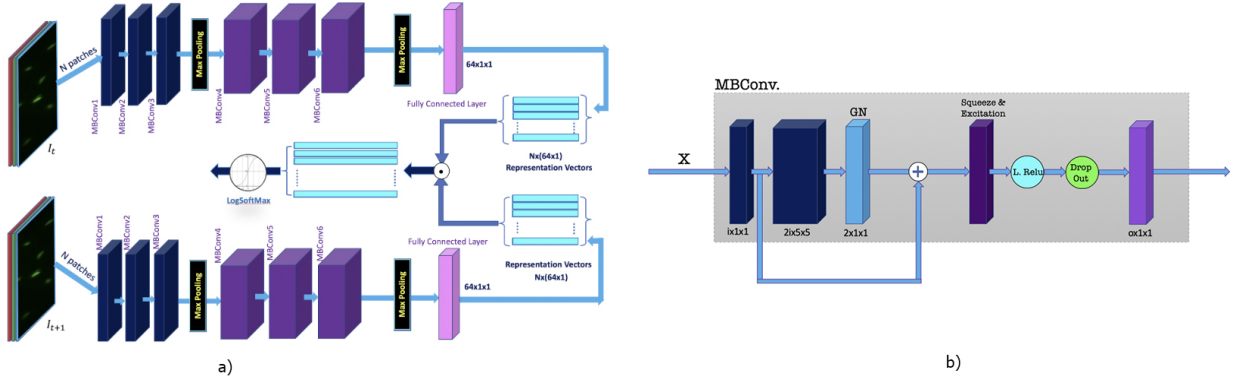


Figure 4.1: (a) The architecture of our contrastive network applied to two consecutive frames. (b) The internals of an MBConv layer.

convolutional neural network, whose final output layer produces a d -dimensional vector, i.e., $v_i = f_\theta(x_i)$, where θ represents the parameters of the neural network f .

The goal of contrastive learning is to bring similar vectors closer together while pushing dissimilar ones farther apart. To achieve this, we need to sample vectors that should be similar and others that should be dissimilar. We use the observation that our videos are Nyquist sampled, i.e., the sampling rate in our videos is high relative to the frequency of the recorded motion. This implies that consecutive frames differ only slightly in content. Therefore, a patch x_i from the same location in two consecutive frames I_t and I_{t+1} will most likely be similar, and this forms the basis for sampling positive examples.

For negative examples, however, we sample random patches from both the current frame and the next frame. Even though these random patches might contain objects visually similar to the current patch, we assume that the corresponding patch from the next frame will be the most similar to the current patch and should thus be coupled positively above any other pairing. We set a ratio of $m : 1$ for negative to positive samples to contrast with the vectors from the current frame.

4.3.3 Datasets

To evaluate the performance of our proposed segmentation method, we utilize a diverse set of biomedical video datasets. By incorporating datasets with a wide range of cell shapes, sizes, and motility patterns, we aim to assess the generalizability of our method across different biological structures and imaging conditions. By applying our method to these datasets, we also aim to evaluate its ability to handle complex, diffuse, or punctate patterns.

Table 4.1: Cell Tracking Challenge (CTC) 2D Datasets

Dataset Name	Modality	Cell Type
BF-C2DL-HSC	Brightfield (BF)	Mouse hematopoietic stem cells
BF-C2DL-MuSC	Brightfield (BF)	Mouse muscle stem cells
DIC-C2DH-HeLa	DIC	HeLa cells on a flat glass
Fluo-C2DL-Huh7	Fluorescence (Fluo)	Human hepatocarcinoma-derived cells
Fluo-C2DL-MSC	Fluorescence (Fluo)	Rat mesenchymal stem cells
Fluo-N2DH-GOWT1	Fluorescence (Fluo)	GFP-GOWT1 mouse stem cells
Fluo-N2DL-HeLa	Fluorescence (Fluo)	HeLa cells expressing H2b-GFP
PhC-C2DH-U373	Phase Contrast (PhC)	Glioblastoma-astrocytoma U373 cells
PhC-C2DL-PSC	Phase Contrast (PhC)	Pancreatic stem cells
Fluo-N2DH-SIM+	Fluorescence (Fluo)	Simulated nuclei of HL60 cells

Cell Tracking Challenge Datasets

We use all available 2D datasets from the Cell Tracking Challenge (CTC) [68]. These datasets include various cell types and imaging modalities, such as fluorescence and phase-contrast microscopy images. They cover a range of biological structures and provide a diverse testbed for evaluating the performance of segmentation methods across different imaging conditions.

4.3.4 Training process

Each iteration of the training, we construct the matrix $R^{n \times d}$ which is the set of patches of an image I_t after passing them through the representation network where column i represents $v_i = f_\theta(x_i)$. To represent the similarity with all the negative and positive samples, we construct the matrix $M^{n \times (m+1)}$ where each column is the dot product between the matrix R with a matrix $Q^{n \times d}$ of random patches sampled randomly from the $2N$ available patches at hand from the current I_t and next I_{t+1} frames, except for the last column, the column of positive patches, where the matrix Q is set to be the vectors of the next patches of the matrix R from the next frame I_{t+1} . We also transform the next-frame positive patches by flipping them horizontally and vertically each with probability 0.5. This is so that the network learns to associate the same texture in different positions and configurations.

4.3.5 Similarity and loss

We choose the cosine similarity between vectors as our similarity metric. The vector output of the convolutional network is, therefore, projected onto the $L2$ unit sphere (i.e., normalized), before being used with dot products. For practical numerical stability, though, we use logSoftmax with negative log-likelihood instead of softmax and cross-entropy.

4.3.6 Interactive Segmentation Workflow

Following the contrastive training phase, the model facilitates interactive segmentation of target structures in new images or videos. The workflow supports two modes of user interaction. In the simpler mode, the user initiates segmentation by clicking a single point on an example of their region of interest (ROI). The image patch centered at this point serves as a positive anchor; its learned embedding vector is compared against the embeddings of all patches across the image using cosine similarity. This process generates a similarity map highlighting regions visually consistent with the anchor patch, which the user subsequently thresholds to produce a binary segmentation mask.

For scenarios requiring finer control or dealing with more complex backgrounds, the user can employ a more complex interaction mode. This involves selecting multiple positive points within the ROI (via clicks) and specifying multiple negative points representing background or unwanted structures (e.g., via Shift + click). The system then calculates an average embedding vector for the positive anchor points and another for the negative points. The final similarity map is computed based on similarity to the average positive embedding while maximizing dissimilarity to the average negative embedding (conceptually achieved by subtracting the negative embedding influence from the positive one). The user then interactively selects a threshold value for this refined similarity map to generate the final segmentation.

4.4 Results and Discussion

We evaluate our method on a subset of 2D datasets from the CTC. The CTC offers a diverse array of 2D and 3D time-lapse microscopy datasets, each capturing unique biological specimens under various imaging modalities. Table 4.1 contains an overview of these datasets, detailing the organisms studied, imaging techniques employed, and acquisition specifics.

For each dataset, or part of dataset, we leave out 20% of the data as a testing portion, and of the remaining 80%, we take 70% of it for training, and 30% for validation. We use the loss on the validation to choose the best model, and report the dice coefficient using the best trained model on the testing portion. In each iteration we sample 1024 patches within the input image, and construct the matrix with the number of negative samples $m = 9$, and the size of representation vector $d = 64$. As noted before, the contrastive loss only minimizes a lower bound on the error, so the training error of the negative log likelihood loss never goes down to 0. We train to 50 epochs for each part of the dataset and use the Adam optimizer as well with the same $10e^{-3}$ learning rate. To generate masks we take user’s input in the form of at least one point indicating the coordinates of the object of interest. These coordinates represent the center of the patch whose representation vector will be the anchor to compare against. We sweep the entire image with patches of size 15×15 and stride of 1, generating a representation vector per each pixel in the image, and report the dot product of these vectors and the anchor vector. We use reflective padding instead of zero padding. Finally, the user can select a suitable threshold to binarize the raw mask into the final mask.

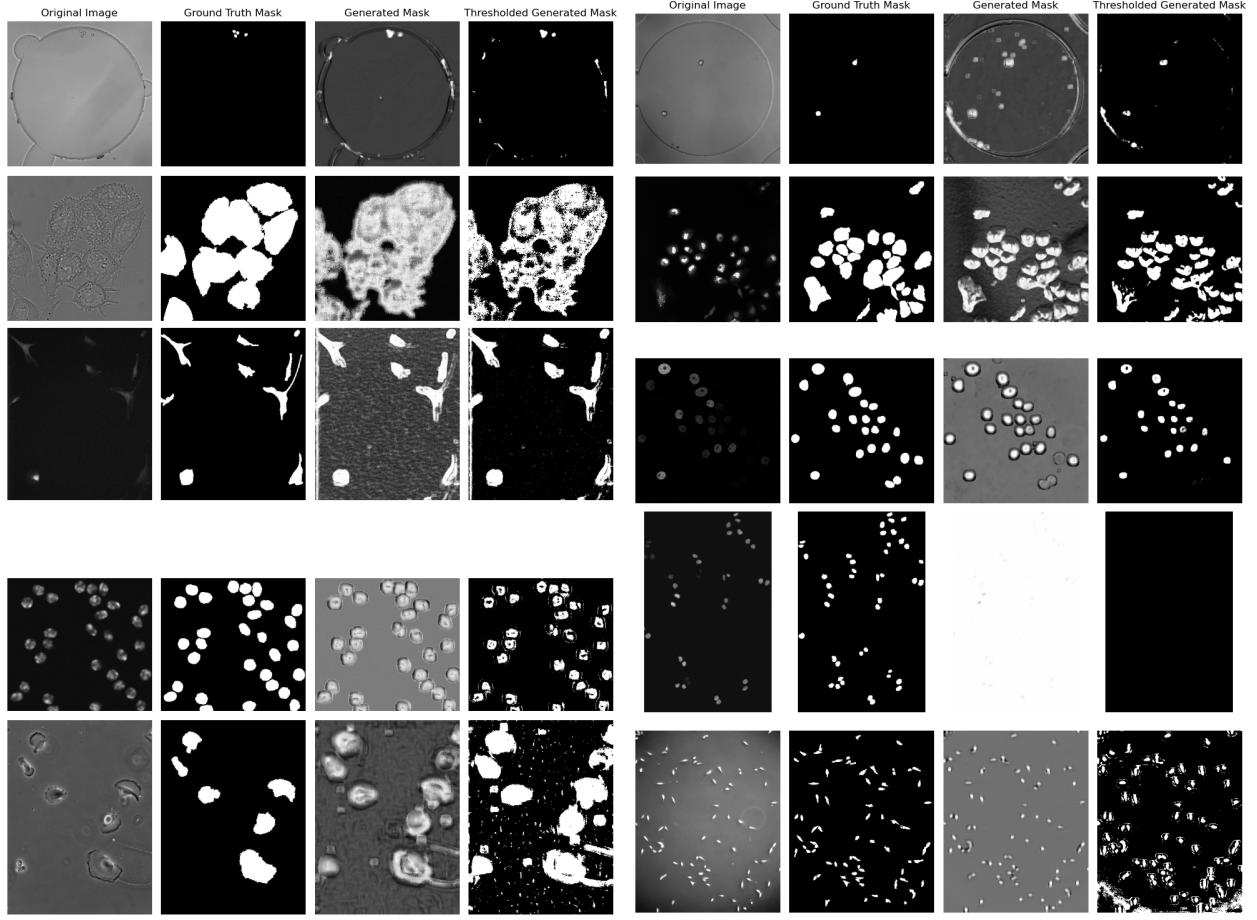


Figure 4.2: Visual comparison of segmentation performance across different datasets. The variation in performance across datasets indicates the challenges caused by different imaging modalities and cell types.

For datasets BF-C2DL-HSC and BF-C2DL-MuSC, artifacts in the image that looked similar to the ROIs shared similar embedding space thus wrongly identified and highlighted as ROI. A simple pre-processing or post processing could resolve this issue.

Table 4.2 shows a mix of strong and moderate results for dice coefficients and precision. The BF-C2DL-HSC and BF-C2DL-MuSC datasets exhibit the weakest performance, with Dice scores of 0.341 and 0.261, respectively. This poor performance is due to the artifacts apparent in the images, which share a strong resemblance with the target stem cells in texture and intensity. These artifacts and shadows within the hydrogel environment likely contribute to false positives and inconsistent mask predictions. The DIC-C2DH-HeLa dataset achieves a moderate Dice score of 0.711, showing a reasonable ability to capture cell structures. However, the fine-grained details of the HeLa cells in differential interference contrast (DIC)

Table 4.2: Dice coefficients and intersection-over-union (IoU) scores for the CTC datasets.

Dataset Name	IoU (JI)	Dice	CTC Benchmark JI
Fluo-N2DH-GOWT1	0.815	0.8971	0.938
Fluo-N2DL-HeLa	0.487	0.655	0.923
PhC-C2DL-PSC	0.736	0.847	0.756
Fluo-C2DL-Huh7	0.617	0.762	0.811
Fluo-N2DH-SIM+	0.786	0.7404	0.832
BF-C2DL-HSC	0.206	0.341	0.855
BF-C2DL-MuSC	0.15	0.261	0.784
DIC-C2DH-HeLa	0.551	0.711	0.877
Fluo-C2DL-MSC	0.419	0.591	0.687
PhC-C2DH-U373	0.342	0.51	0.931

JI = Jaccard Index, IoU = Intersection over Union. The CTC benchmark ¹ JI is the best performing algorithm on the CTC datasets in CTC's website.

imaging pose difficulties in maintaining sharp boundary delineation, leading to a loss in segmentation accuracy. The top performing algorithms on the 2D CTC datasets are shown in Table 4.3 for comparison.

Table 4.3: Top-Performing Jaccard Index Scores on 2D CTC Datasets.

Dataset/Metric	Jaccard Index
BF-C2DL-HSC	0.895
BF-C2DL-MuSC	0.784
DIC-C2DH-HeLa	0.871
Fluo-C2DL-Huh7	0.811
Fluo-C2DL-MSC	0.687
Fluo-N2DH-GOWT1	0.938
Fluo-N2DL-HeLa	0.913
PhC-C2DH-U373	0.931
PhC-C2DL-PSC	0.756

For the Fluo-C2DL-MSC dataset, the Dice coefficient of 0.591 indicates moderate segmentation quality. The elongated morphology of mesenchymal stem cells complicates boundary definitions, leading to thresholding artifacts. The PhC-C2DH-U373 dataset, with a Dice coefficient of 0.510, shows lower performance due to halo effects in phase contrast imaging, which interfere with precise boundary extraction and introduce noise.

The moderate performance on the DIC-C2DH-HeLa dataset (Dice 0.711) can be further understood by considering the large size and complex internal structure of these cells relative to the fixed patch size (15x15 pixels) used by our model. Since contrastive learning operates by comparing the local texture within these small patches, an anchor point selected within a large cell represents only a specific internal feature. Consequently, the resulting segmentation tends to highlight only those regions across the image sharing

that particular local texture, often leading to incomplete masks that capture parts of the cell rather than its entirety.

Challenges were also encountered with the Fluo-N2DL-HeLa dataset (Dice 0.655), primarily stemming from the very low contrast between the labeled cells and the image background. This minimal visual distinction makes it difficult for the contrastive learning approach, which relies on differentiating similar and dissimilar patches, to effectively separate cell regions from the background, thus hindering accurate boundary delineation.

Overall, the results suggest that datasets with clear and well-defined fluorescence-stained boundaries (such as GOWT1) tend to perform best, while datasets relying on phase contrast, brightfield, or DIC imaging suffer from boundary inconsistencies and intensity variations that complicate segmentation.

4.4.1 Potential Improvements

Beyond threshold adjustments, a number of strategies could further improve performance on challenging datasets:

- **Use Multi-scale or Pyramid Features.** Incorporating a multi-scale approach (e.g., learning features from multiple patch sizes) can help capture both the global context of larger cells and the subtle boundaries in smaller or low-contrast targets.
- **Refine Negative Sampling.** Randomly sampling negatives can push away genuinely similar patches. A more stratified or context-aware sampling strategy can reduce such conflicts and help preserve true positives.
- **Domain-specific Augmentations.** For brightfield or DIC images, augmentations such as slight intensity inversions or custom brightness manipulations can help the model learn invariant features against heterogeneous illumination and shadows.
- **Smaller Patches with Additional Context.** Reducing the patch size focuses on local cell texture, but simultaneously providing a lower-resolution context channel can keep global cues. This helps avoid confusion from large background areas.
- **Leverage Small Amounts of Label Information.** In practice, adding even sparse or partial labels, even minimal bounding boxes or scribbles, for a few frames can anchor the contrastive model, mitigating background confusion in tough domains.

These enhancements can bolster the robustness of contrastive embeddings, especially in complex imaging conditions such as brightfield, DIC, or phase contrast, where artifacts and shadows closely resemble cellular structures.

4.5 Conclusion and Final Remarks

We have introduced a method for object segmentation using contrastive coding, requiring minimal user input to select the object of interest. By enforcing similarity between temporally adjacent patches and differentiating dissimilar ones, the model learns embeddings that enable segmentation without the need for labeled datasets. Our approach generates visually plausible masks and demonstrates good results in some datasets, although it achieves only moderate performance in others. We discuss the reasons for these varying results and emphasize the novelty of our method. Additionally, we provide a GUI tool to assist users in marking the object of interest and setting the appropriate threshold for the entire video.

While our results are encouraging, there are several avenues for future work. First, refining the segmentation boundaries through post-processing or boundary-focused contrastive objectives could help address residual errors in challenging datasets. Second, extending this framework to three-dimensional or volumetric time-series data would further increase its applicability to advanced imaging techniques. In conclusion, our contrastive learning-based approach offers a scalable and practical alternative to fully supervised or foundation-model-driven segmentation pipelines, enabling segmentation of diverse biomedical structures with zero annotation.

CHAPTER 5

TOWARD A FOUNDATION MODEL FOR BIOMEDICAL IMAGE SEGMENTATION

5.1 Introduction

Biomedical image segmentation is pivotal across a diverse array of medical and biological applications, from diagnostic imaging to cellular analysis. While supervised segmentation methods, especially convolutional neural networks (CNNs), have achieved remarkable accuracy, their reliance on large, annotated datasets severely limits their generalizability and transferability, particularly in biomedical contexts where labeled data is scarce and costly to obtain. Conversely, unsupervised methods, despite being simpler and more generalizable, often fall short in segmentation precision and robustness, thus failing to meet the accuracy requirements of clinical and research settings.

Recently, the emergence of Foundation Models (FMs) and the Segment Anything Model (SAM) offers a compelling new direction for addressing these limitations. SAM, introduced by Kirillov et al. [49], marked a significant advancement by achieving impressive zero-shot segmentation capabilities across varied image domains, relying only on minimal user prompts such as points or bounding boxes. Its successor, SAM 2, further extends these capabilities into video segmentation, leveraging advanced architectures to improve accuracy and interaction efficiency [79]. These models are trained on large and diverse datasets, enabling them to generalize effectively across multiple segmentation tasks without extensive domain-specific training. Despite their potential, the application of general-purpose SAM models to biomedical imaging presents unique challenges. The complex and nuanced nature of biomedical images characterized by varying imaging modalities, structures, textures, and noise levels means that models trained primarily on general-domain images may struggle to achieve desirable segmentation accuracy [69], [71]. Recognizing this, adaptations of SAM tailored specifically to biomedical contexts have emerged. Models such as MedSAM [65] and MediViSTA-SAM [47] demonstrate the feasibility of adapting SAM to medical imaging and video analysis, showing promising results that often surpass specialized, modality-specific models in robustness and accuracy.

Alongside spatial segmentation capabilities, the integration of vision-language models like BiomedCLIP [111] into SAM workflows has gained considerable attention. BiomedCLIP, trained on extensive biomedical image-text pairs, provides robust multimodal embeddings that bridge textual and visual domains, enabling powerful text-driven segmentation. Frameworks such as MedCLIP-SAMv2 exemplify this integration, demonstrating how textual prompts can effectively guide precise segmentation tasks, from identifying tumors in medical scans to delineating specific cellular structures in microscopy images. However, the integration of vision-language models with SAM introduces additional layers of complexity and stochasticity. Variability in outputs stemming from factors such as the fine-tuning process of models like BiomedCLIP, differences in textual prompts, or randomness in inference strategies can undermine reproducibility, a critical factor in biomedical research and clinical applications. Thus, understanding and mitigating this stochasticity is paramount for the practical adoption and reliability of these models.

This chapter aims to comprehensively explore and address these challenges. We focus specifically on evaluating and enhancing the reproducibility of integrated vision-language segmentation models, investigating how different factors—including fine-tuning strategies, prompt engineering, and inference methodologies—influence variability in segmentation outcomes. Our goal is to develop methodologies that standardize and optimize these variables, ensuring SAM-based models are not only accurate and versatile but also reliably reproducible in biomedical contexts.

5.2 Background

SAM [49] introduced a groundbreaking promptable segmentation approach, trained on a vast dataset (SA-1B), featuring over one billion masks. Its architecture comprises three components: a powerful Vision Transformer (ViT) image encoder, a flexible prompt encoder handling points, boxes, and text, and a lightweight mask decoder to produce segmentation masks. SAM’s notable innovation lies in its zero-shot capabilities achieved through prompt engineering, enabling it to generalize well across various segmentation tasks, even without task-specific fine-tuning. SAM-2 [79] extends SAM’s capabilities to video data, incorporating a memory attention mechanism that retains information from previous frames, thereby significantly enhancing segmentation accuracy and interaction efficiency. SAM-2 utilizes a hierarchical transformer architecture (Hiera) [3], [82] pre-trained with masked autoencoders (MAE) [27], making it highly effective for real-time segmentation tasks across images and videos. This memory-enhanced architecture allows SAM-2 to iteratively refine masks, leading to considerable improvements in temporal segmentation consistency. A comprehensive survey titled "Foundation Models for Biomedical Image Segmentation" [35] underscores the transformative potential of SAM, summarizing over 100 studies that have successfully adapted SAM to a wide range of biomedical datasets. The survey highlights SAM’s strong zero-shot capabilities and outlines various domain-specific tuning methods and data scarcity challenges that have driven innovation in biomedical segmentation.

The success of the SAM in general-domain image segmentation has inspired adaptations and methodological enhancements aimed at tailoring its capabilities specifically for biomedical applications, addressing unique challenges associated with medical image segmentation. MedSAM [65] and other models, like

BioSAM-2 [106], have demonstrated the necessity of domain-specific fine-tuning for achieving clinical-grade accuracy. BioSAM-2, particularly designed for biomedical segmentation, has optimized SAM-2 with medical domain-specific data and additional memory mechanisms for improved performance across diverse biomedical imaging modalities. Medical SAM Adapter (Med-SA) [102] introduced adaptation modules such as Space-Depth Transpose (SD-Trans) and Hyper-Prompting Adapter (HyP-Adpt), which enhance SAM’s performance on medical images through minimal yet strategic parameter adjustments. These modules have shown significant improvements over traditional segmentation methods by efficiently incorporating medical domain knowledge.

In the realm of prompt learning and auto-prompting the Segment Any Cell (SAC) [71] framework leveraged auto-prompting and fine-tuning methods, using Low-Rank Adaptation (LoRA) [31], to automatically generate effective prompts for nuclei segmentation. This method reduced manual intervention and improved segmentation accuracy in microscopic imaging scenarios. SSPrompt [32] optimized SAM’s spatial and semantic prompts directly within its embedding space, enhancing its generalization capabilities across complex segmentation tasks. The Segment and Caption Anything [33] model enriched SAM’s semantic understanding capabilities by integrating a query-based feature mixer, improving semantic precision and enabling the model to provide meaningful regional captions, thus enhancing segmentation results through better semantic contextualization.

The support for textual prompts in the original SAM is relatively limited and experimental compared to spatial (points, boxes, masks) prompts. Building upon the advancements in adapting SAM for biomedical tasks, recent research has increasingly focused on incorporating text prompts and integrating vision-language models to further enhance segmentation precision and semantic interpretability in medical imaging. Models like BiomedCLIP [111] and adaptations such as MedCLIP-SAMv2 [50], [51] underscore the importance of text-driven segmentation approaches, leveraging extensive biomedical image-text pairs to provide robust multimodal embeddings. This integration enables powerful and precise segmentation guided by textual descriptions, thus bridging visual and textual biomedical data effectively. The EVF-SAM [113] model exemplifies the integration of early vision-language fusion. It incorporates an early fusion mechanism, significantly outperforming late fusion models by enhancing text-to-image attention, which is critical for accurate segmentation guided by referring expressions.

Polyp-SAM++ [2] demonstrated the effectiveness of detailed textual prompts specifically for colorectal polyp segmentation, showing how text guidance could substantially improve the segmentation accuracy and robustness of SAM, particularly in clinically relevant contexts. Hi-SAM [107] extended SAM’s capabilities to hierarchical text segmentation, including pixel-level text, word, text-line, and paragraph segmentation, thus enabling more structured and detailed biomedical image analyses, crucial for applications like pathology slide examination. PROMISE [59] and similar models have adapted SAM to 3D biomedical segmentation, introducing lightweight adapters for depth-related spatial context and achieving superior performance in tumor segmentation tasks by effectively combining textual prompts with depth-awareness.

The Segment Anything with Text prompts (SAT) [114] model, trained on an extensive dataset comprising over 22,000 medical scans and nearly 500 anatomical classes, exemplifies a universal segmentation

framework that integrates extensive medical terminologies as textual prompts. This approach emphasizes the utility of incorporating domain-specific knowledge directly into the model training, significantly improving segmentation performance across diverse medical imaging tasks.

When using text-to-segmentation pipelines, segmentation results should ideally be deterministic given the same input. However, randomness can creep in through various stages, both at inference and during training. There is growing interest in modeling the stochasticity and uncertainty inherent in medical image segmentation. In practice, what constitutes the "correct" segmentation can be ambiguous – different experts may trace slightly different boundaries for the same lesion, especially in low-contrast or complex cases [78]. Models like SAM produce one deterministic mask per prompt, which doesn't capture this ambiguity, however, this ambiguity is maximised when incorporating text-to-segment pipelines, since the text-image input of the CLIP should be transformed into spacial points to be fed to SAM as input prompts. In MedCLIP-SAMv2 [50] this step is done with the help of extracting attention maps of BiomedCLIP given a text-image pair. This saliency map highlights the locations of interest in the image which are then used to select points or bounding boxes for SAM.

Our methodology builds upon the principles of MedCLIP-SAMv2, integrating the BiomedCLIP vision-language model with SAM to improve segmentation accuracy and reproducibility in biomedical applications. Initially, we evaluate the baseline capability of BiomedCLIP to differentiate various biological structures, emphasizing its performance on complex and previously unseen entities such as ciliary regions. Subsequently, we explore the impact of fine-tuning BiomedCLIP, examining how variations in fine-tuning parameters influence its effectiveness. Lastly, we assess multiple strategies for selecting optimal spatial prompts as inputs for SAM, aiming to identify methods that consistently yield accurate segmentation outcomes.

5.3 Methodology

This chapter outlines a systematic approach developed to enhance segmentation accuracy and reproducibility in biomedical imaging, specifically targeting the segmentation of ciliary regions in nasal epithelial biopsy videos. Leveraging the BiomedCLIP vision-language model integrated with the SAM, the methodology involves fine-tuning BiomedCLIP on a specialized dataset containing annotated videos of nasal epithelial biopsies. These annotations distinctly mark cell bodies and associated ciliary regions, which vary significantly in visibility—ranging from clearly delineated, easily identifiable structures to overlapping and out-of-focus regions challenging even to expert human annotators. The methodology rigorously explores how different fine-tuning parameters, textual prompts, and image pre-processing strategies (masked versus raw) influence BiomedCLIP's performance. Subsequently, the trained BiomedCLIP generates predictive heatmaps on previously unseen data, serving as a basis for strategically selecting spatial prompts for SAM segmentation. Finally, this research investigates various prompt selection strategies to determine optimal methods for ensuring consistent, accurate, and reproducible segmentation results.

5.3.1 Datasets

Cilia Dataset

Our principal dataset comprises 681 microscopy videos of nasal epithelial biopsies, among which 325 videos have detailed annotations. Each video depicts epithelial cells exhibiting either normal motile cilia, dyskinetic (immotile) cilia. Annotated masks explicitly identify three types of regions: cell bodies, clearly visible cilia structures, and overlapping, hard-to-detect cilia. Visible cilia typically extend beyond cell boundaries, appearing clearly against a blank background, thus facilitating their straightforward identification in static frames. Conversely, overlapping cilia structures, often oriented vertically toward the microscope lens or appearing sparse and out of focus, pose significant detection challenges, even to human annotators. Such challenging structures usually require observing subtle rhythmic patterns across video frames for confident identification.

For simplicity and consistency within our methodology, both visible and overlapping ciliary regions were grouped into a single class termed "ciliary structure." To evaluate the effectiveness of fine-tuning BiomedCLIP and subsequent segmentation using SAM, the annotated cilia dataset was split into training and testing subsets following a 70/30 ratio.

The Cell Tracking Challenge Dataset [68]:

We employed selected datasets from the Cell Tracking Challenge (CTC), which offers a comprehensive collection of 2D and 3D time-lapse microscopy images, each representing diverse biological organisms and imaging modalities. These datasets cover a broad range of specimens, including human, mouse, rat, *Caenorhabditis elegans*, *Drosophila melanogaster*, and others, captured using modalities such as Brightfield, Differential Interference Contrast (DIC), and Fluorescence microscopy (detailed in Table 2.1). Given the relative uniformity of image slices within individual datasets, a small representative sample from each was sufficient for benchmarking BiomedCLIP's inherent segmentation capabilities prior to fine-tuning.

BiomedCLIP Model

BiomedCLIP, a contrastive vision-language model tailored specifically for biomedical domains, integrates effectively with segmentation frameworks like SAM, exemplified by models such as MedCLIP-SAMv2. Trained on 15 million biomedical image-text pairs from PubMed Central, BiomedCLIP leverages a vision transformer (ViT) for image encoding and PubMedBERT [23] for text encoding. Through contrastive learning, it aligns related images and textual descriptions into a joint embedding space. Unlike general-purpose models like CLIP [76], BiomedCLIP captures nuanced, domain-specific features, making it uniquely suited for precise biomedical image analysis tasks, including saliency-driven segmentation. Figure 5.1 shows the off the shelf performance of BiomedCLIP over 2D datasets in CTC. The results fall

significantly short of the performance achieved by state-of-the-art models in the CTC, as shown in Table 2.1¹.

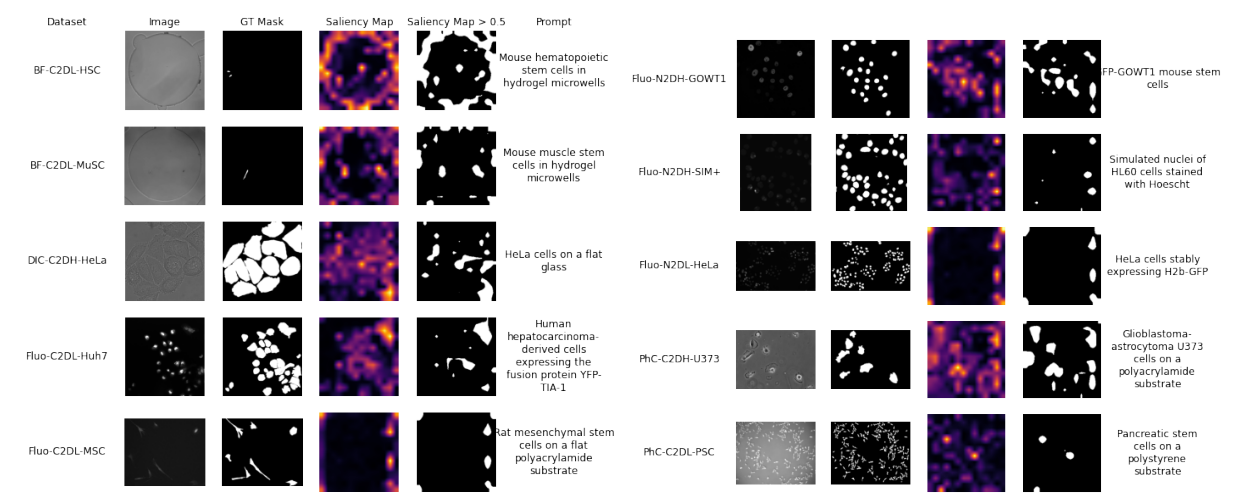


Figure 5.1: BiomedCLIP’s out-of-the-box performance over CTC’s 2D datasets

Fine-Tuning Strategies

To investigate the impact of different fine-tuning parameters on BiomedCLIP’s performance for cilia segmentation, we systematically varied several key factors. These included adjusting the number of training epochs, comparing fine-tuning on raw images versus masked images (isolating cell bodies and ciliary structures), utilizing varying lengths of textual descriptions as noted in Table 5.5 (short versus detailed versus randomized detailed annotations), and experimenting with different configurations of loss function parameters. The parameter variations are summarized in Table 5.1.

Table 5.1: Fine-Tuning Parameter Variations

Parameter	Variation Descriptions
Number of Epochs	2 epochs vs. 32 epochs
Input Images	Raw images vs. Masked images
Textual Descriptions	Concise vs. Detailed Vs. Randomized detailed descriptions
Loss Function Parameters	Default settings vs. Adjusted weighting schemes

The fixed parameters for fine-tuning included the pretrained BiomedCLIP model, a batch size of 32, learning rate of 1×10^{-3} , weight decay of 0.1, training duration of 32 epochs, and the DHN-NCE loss introduced in MedCLIP-SAMv2. Model checkpoints were saved after each epoch.

¹More information about the top-performing algorithms can be found <https://celltrackingchallenge.net/latest-csb-results/>

DHN-NCE Loss Function

The Decoupled Hard Negative Noise Contrastive Estimation (DHN-NCE) loss is designed to improve contrastive learning by decoupling positive samples from the denominator and introducing hard negative sampling. The loss function consists of two terms: one for image-to-text learning and another for text-to-image learning. Table 5.2 describes the effects of increasing or decreasing each of the loss-function parameters.

Loss Definition

The overall DHN-NCE loss is defined as:

$$L_{\text{DHN-NCE}} = L_{v \rightarrow t} + L_{t \rightarrow v} \quad (5.3.1)$$

Each term is computed as follows:

$$L_{v \rightarrow t} = - \sum_{i=1}^B \frac{I_{p,i} T_{p,i}^\top}{\tau} + \sum_{i=1}^B \log \left(\sum_{j \neq i} e^{I_{p,i} T_{p,j}^\top / \tau} W_{v \rightarrow t} \right) \quad (5.3.2)$$

$$L_{t \rightarrow v} = - \sum_{i=1}^B \frac{T_{p,i} I_{p,i}^\top}{\tau} + \sum_{i=1}^B \log \left(\sum_{j \neq i} e^{T_{p,i} I_{p,j}^\top / \tau} W_{t \rightarrow v} \right) \quad (5.3.3)$$

where:

- $I_{p,i}$ and $T_{p,i}$ are the normalized image and text features.
- B is the batch size.
- τ is the temperature parameter controlling the sharpness of the distribution.
- $W_{v \rightarrow t}$ and $W_{t \rightarrow v}$ are hardness weights for negative samples.

Hardness Weighting Factors

The hardness weighting factors are defined as:

$$W_{v \rightarrow t} = (B - 1) \times \frac{e^{\beta_1 I_{p,i} T_{p,j}^\top / \tau}}{\sum_{k \neq i} e^{\beta_1 I_{p,i} T_{p,k}^\top / \tau}} \quad (5.3.4)$$

$$W_{t \rightarrow v} = (B - 1) \times \frac{e^{\beta_2 T_{p,i} I_{p,j}^\top / \tau}}{\sum_{k \neq i} e^{\beta_2 T_{p,i} I_{p,k}^\top / \tau}} \quad (5.3.5)$$

Table 5.2: Effects of parameter changes on DHN-NCE loss.

Parameter	Effect of Increasing	Effect of Decreasing
Temperature (τ)	Smoother similarity distribution	Sharper contrast, possible instability
Hardness (β_1, β_2)	Emphasizes difficult negatives	Less focus on hard negatives
Positive Weight (α)	More weight on positives, less contrast	Stronger negative differentiation

Evaluation of BiomedCLIP Predictions

To quantitatively assess the accuracy of the generated heatmaps, a thresholding method was required prior to computing evaluation metrics such as Dice coefficient and Intersection over Union (IoU) against ground-truth masks. Our experiments indicated that the threshold selection critically influences the evaluation outcomes, and the optimal threshold value differs markedly across various fine-tuning configurations. Specifically, models with minimal or no fine-tuning necessitated relatively lower thresholds, whereas models subjected to extensive fine-tuning required comparatively higher thresholds to achieve optimal segmentation performance.

5.4 Results and Discussion

To establish a baseline performance for localizing ciliary regions in nasal epithelial biopsy images, the off-the-shelf BiomedCLIP model was evaluated without domain-specific fine-tuning. Textual prompts of varying lengths were used during inference: *Short* ("respiratory cilia"), *Medium* ("cilia on nasal epithelial cells"), and *Detailed* ("normal or abnormal cilia in nasal epithelial biopsy...harder to detect"). As qualitatively illustrated in Figure 5.2, the un-tuned model struggled to consistently and accurately identify ciliary structures, often generating saliency maps poorly aligned with the ground truth masks. Quantitative metrics calculated from thresholded saliency maps corroborated these limitations. While average IoU and Dice scores across test samples are presented in Table 5.3, visual inspection and individual sample metrics revealed that IoU scores for specific predictions rarely exceeded 0.01, underscoring the challenges faced by the un-tuned model and motivating the need for fine-tuning. It is also important to note that threshold selection significantly impacts these metrics, adding complexity to direct comparisons based solely on average scores.

Table 5.3: Average IoU and Dice scores of BiomedCLIP models under different training configurations

Model Variant	Average IoU	Average Dice
Un-tuned BiomedCLIP (zero-shot)	0.092	0.0033
Fine-tuned for 32 epochs on full images (overfit)	0.000	0.0000
Fine-tuned $\tau = 0.1, \alpha = 0.9, \beta_1 = 0.95, \beta_2 = 0.05$	0.106	0.0040

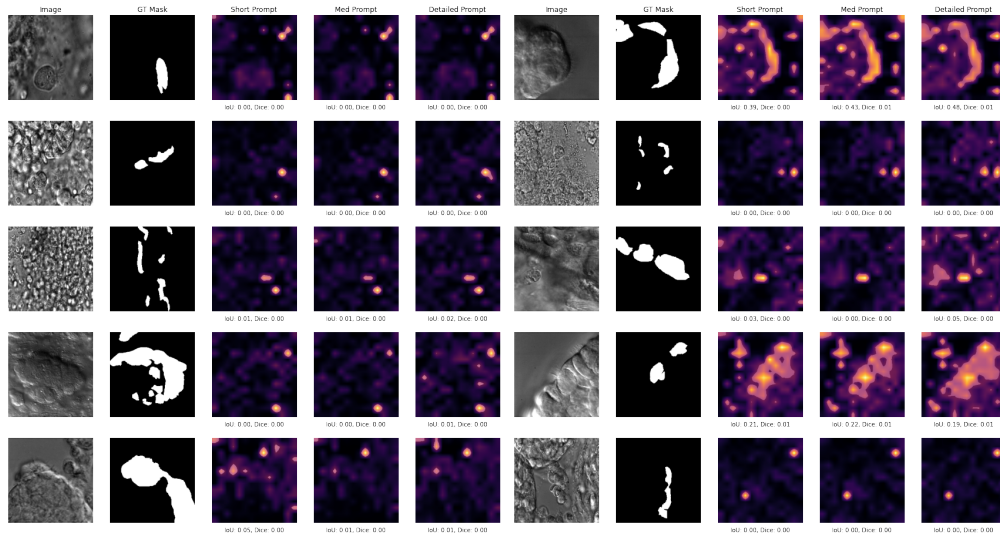


Figure 5.2: Performance of off-the-shelf BiomedCLIP with Short, Medium, and Detailed prompts, demonstrating poor localization without fine-tuning.

5.4.1 Impact of Fine-tuning BiomedCLIP

Our fine-tuning experiments revealed that performance improvements are highly dependent on the specific training conditions. Fine-tuning BiomedCLIP using full-sized, raw images proved detrimental, severely impairing the model’s localization capabilities regardless of the number of training epochs, likely due to the model failing to focus on relevant features amidst complex backgrounds. Conversely, fine-tuning using masked images, which isolate the cell bodies and ciliary structures, consistently improved localization accuracy by focusing the model’s learning process. However, training beyond an optimal number of epochs (empirically found to be around 5 epochs for configurations using masked images) did not enhance performance further. Instead, extended training introduced overfitting and degraded generalizability, as indicated by the increasingly sparse and overly specialized activation maps shown for a 32-epoch run in Figure 5.3.

A quantitative assessment using per-pixel confusion matrix metrics (Table 5.4) further elucidates the impact of fine-tuning. Compared to the untuned model, the fine-tuned version (using the configuration specified in Table 5.3) demonstrated improved sensitivity, achieving higher average True Positive (TP) counts and lower average False Negative (FN) counts across Short, Medium, and Detailed prompts. This confirms the fine-tuned model’s enhanced ability to correctly identify actual positive pixels (cilia). However, this gain in sensitivity was accompanied by a significant reduction in specificity, evidenced by the substantial increase in average False Positive (FP) counts and the corresponding decrease in average True Negative (TN) counts. This indicates the fine-tuned model became more liberal in its classifications, incorrectly labeling a larger number of background pixels as positive. Notably, the performance variation

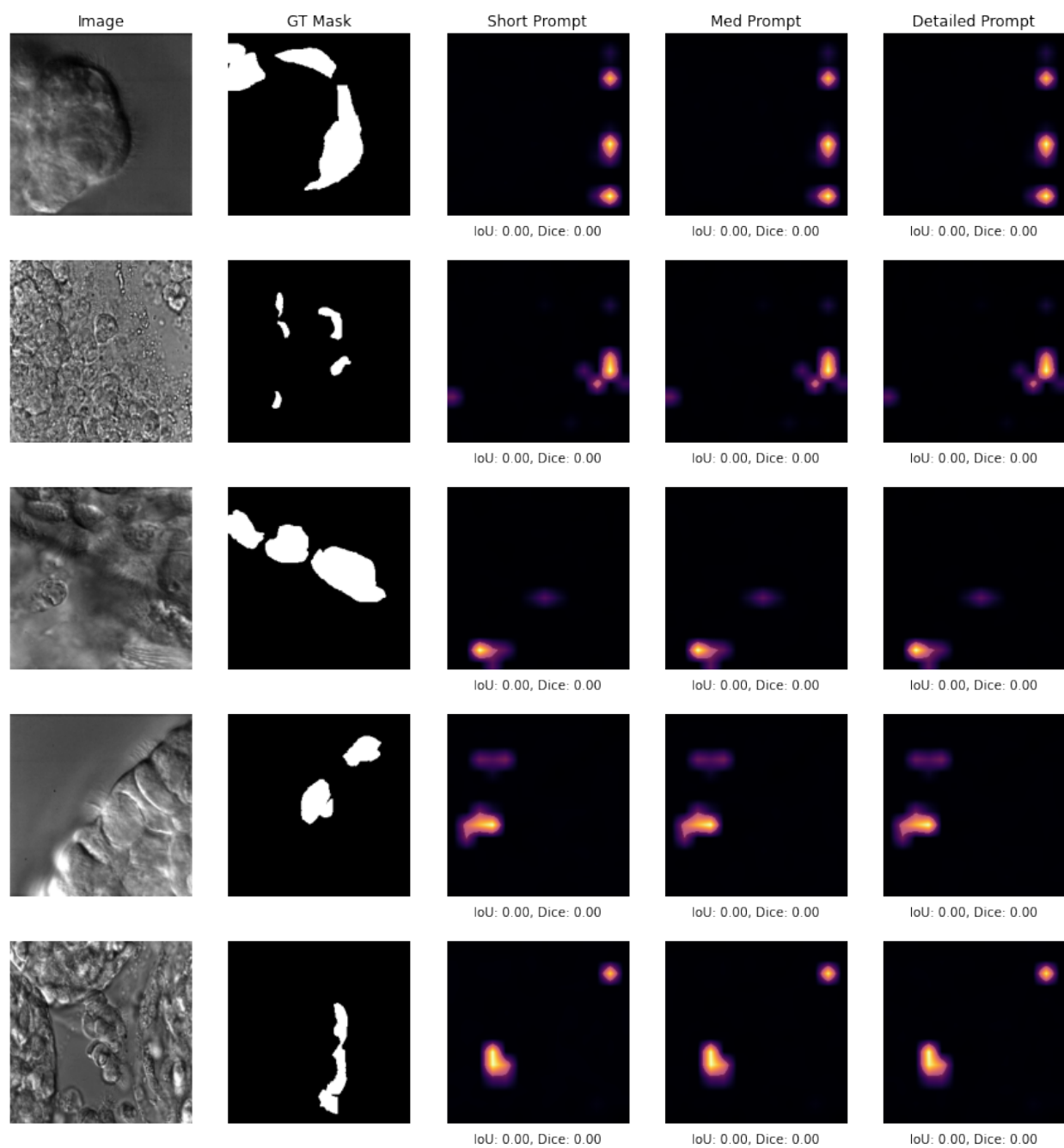


Figure 5.3: Localization results after fine-tuning BiomedCLIP for 32 epochs on full images without masks, showing incorrect and overly specialized localization.

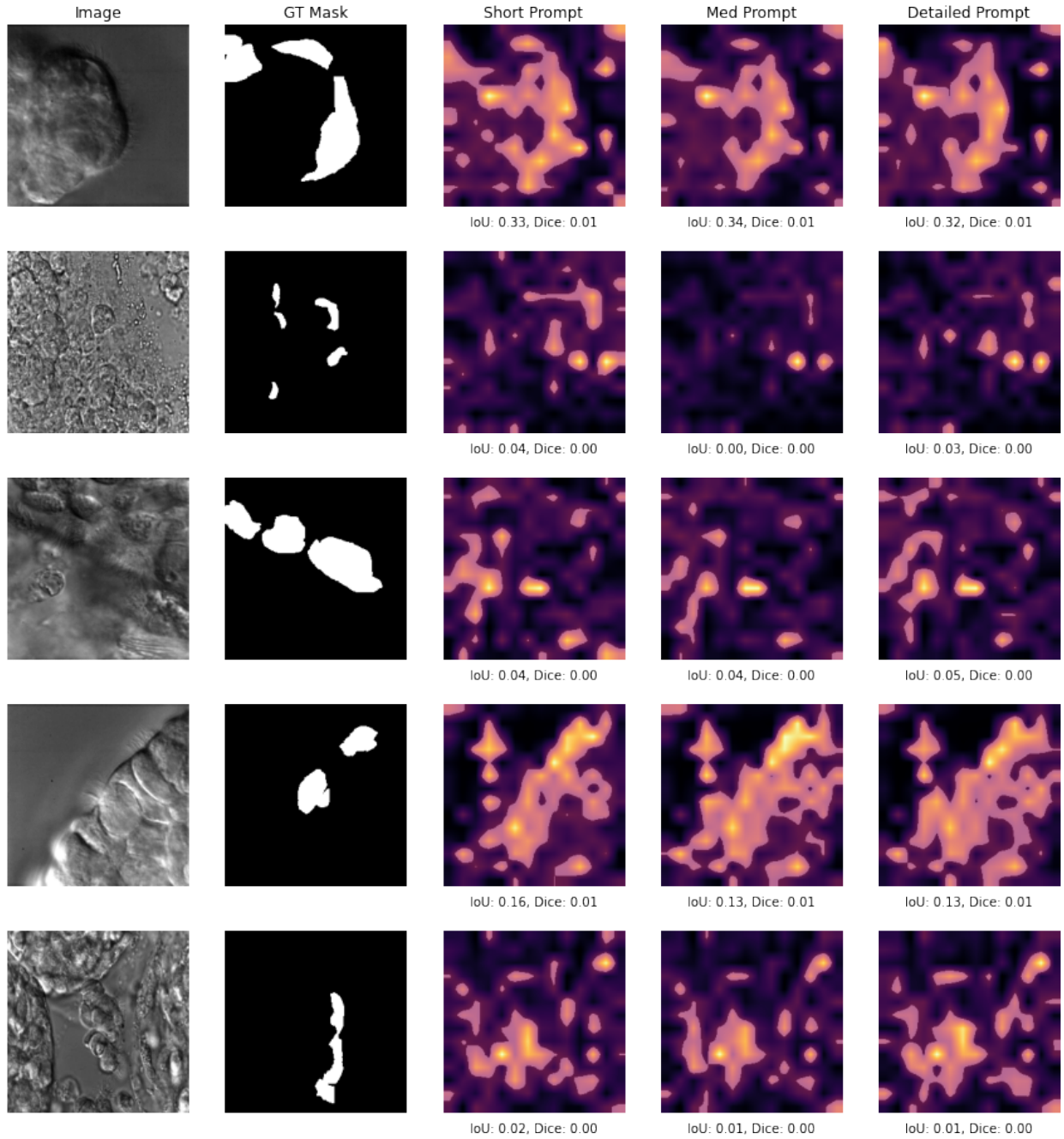


Figure 5.4: Impact of adjusted hardness parameters ($\beta_1 = 0.65, \beta_2 = 0.65$) during fine-tuning, showing improved localization diversity but potentially higher false-positive rates compared to baseline fine-tuning.

across different prompt types appeared less pronounced in the fine-tuned model compared to the untuned one based on these metrics.

Table 5.4: Average Per-Pixel Confusion Matrix Metrics (15 Test Samples)

Model Type	Prompt Type	Avg. TP	Avg. TN	Avg. FP	Avg. FN
Untuned	Short	1370.60	43288.20	2802.40	2714.80
	Medium	1203.40	43001.00	3089.60	2882.00
	Detailed	1642.80	42078.60	4012.00	2442.60
Fine-tuned	Short	1839.80	37405.00	8685.60	2245.60
	Medium	1669.80	37407.60	8683.00	2415.60
	Detailed	1713.20	37208.80	8881.80	2372.20

Note: TP = True Positives, TN = True Negatives, FP = False Positives, FN = False Negatives. Values represent average per-pixel counts over 15 test samples. Fine-tuned model corresponds to the configuration: $\tau = 0.1$, $\alpha = 0.9$, $\beta_1 = 0.95$, $\beta_2 = 0.05$, masked images, 5 epochs, trained on long captions.

5.4.2 Effects of Textual Prompts

Table 5.5: Caption Generation Logic Based on Image Content for Fine-Tuning.

Image Content	Short Caption	Long Caption
Masked Cell Body (No Cilia)	Fixed text: nasal epithelial cell	Fixed text: a microscopy image of nasal epithelial biopsy containing epithelial cell
Masked Ciliary Region	Generated based on the source video's label (motile, immotile, or indeterminate): Resulting text is like "normal cilia", "abnormal cilia", or "indeterminate cilia".	Generated based on source video label and mask analysis: Starts with "nasal epithelial biopsy of [state] cilia..." (where <i>state</i> reflects motile/immotile/indeterminate). Continues by describing visibility based on mask characteristics, using phrases like "...exposed and clearly visible protruding out..." for easily detected cilia, or "...overlying the body of the cell and are hard to detect" for obscured cilia.

Note: If both clearly visible and hard-to-detect cilia are present according to the mask analysis, the long caption explicitly mentions this mixed visibility (e.g., "...some cilia protrude... clearly visible, while some other cilia are overlying... hard to detect."). Additionally, to enhance diversity, the exact phrasing and use of synonyms within the descriptive parts of the long captions were randomized during generation.

The nature of textual prompts used during fine-tuning and inference also significantly impacted model performance (details of caption generation logic are in Table 5.5). Our experiments showed that employing short, concise textual prompts generally led to more consistent localization with fewer false-positive results, particularly when the model was also trained using corresponding short captions. Conversely,

utilizing longer, more detailed prompts and captions (as seen in Figure A.3) could introduce ambiguity, sometimes increasing the incidence of both false positives and false negatives. A crucial observation was the importance of consistency between training captions and inference prompts: models trained with brief captions performed best with brief prompts, whereas models trained on detailed captions benefited from detailed prompts during inference.

5.4.3 Hyperparameter Optimization and Stability

Exploration of the DHN-NCE loss hyperparameters revealed nuances in their influence on stability and accuracy. Within a moderate range, variations in the hardness weighting parameters β_1 and β_2 showed minimal impact. However, employing more extreme, asymmetric values (such as $\beta_1 = 0.95$, $\beta_2 = 0.05$, used in the primary fine-tuned configuration reported in Tables 5.3 and 5.4) significantly heightened the model’s sensitivity to differences in textual prompts, potentially leading to instability. Separate experiments indicated that a balanced beta configuration (e.g., $\beta_1 = 0.65$, $\beta_2 = 0.65$), particularly when combined with long captions during training (as illustrated in Figure 5.4), could offer a good trade-off by maximizing detection accuracy while potentially mitigating some false positives compared to the more sensitive configuration. Adjusting the temperature parameter τ also had effects; higher temperatures led to smoother, more blurred similarity distributions, which could compromise localization precision.

5.4.4 Prompt Selection Strategies for SAM

Finally, our analysis underscored the critical importance of how BiomedCLIP-generated saliency maps are translated into spatial prompts for the downstream SAM model. We found that common strategies, such as randomly selecting points solely from the largest detected contour in the thresholded map, were often inadequate, especially for ROIs like cilia that can be sparse and distributed across multiple small regions. Our findings emphasize that more nuanced approaches, which consider multiple salient contours or employ different heuristics based on the map’s characteristics, can significantly enhance segmentation accuracy, particularly for complex images with overlapping or indistinct ROIs.

5.5 Conclusion and Final Remarks

This chapter investigated the application of foundation models for biomedical image segmentation, focusing specifically on optimizing the BiomedCLIP vision-language model to generate high-quality saliency maps intended as input for a downstream Segment Anything Model (SAM). Our work, centered on the challenging task of ciliary region identification, highlighted the limitations of traditional methods and explored how fine-tuning a vision-language model can enhance the initial, prompt-generation stage of a SAM-based segmentation pipeline.

A key finding is the observed performance difference in BiomedCLIP before and after fine-tuning for this specialized task. Our baseline evaluation showed that the off-the-shelf BiomedCLIP model demonstrated limitations in localization accuracy (e.g., low average IoU scores, high false negatives) and produced

inconsistent saliency maps across different prompt types (Table 5.3, Table 5.4). In contrast, strategic fine-tuning resulted in an improvement in BiomedCLIP’s ability to generate stable and accurate saliency maps relevant to the target structures. Although this fine-tuning increased sensitivity (higher True Positives, lower False Negatives), it also reduced specificity, evidenced by the increase in average False Positive counts (Table 5.4). This observed contrast underscores the necessity of domain-specific adaptation for vision-language models when preparing inputs for subsequent segmentation tasks in specialized biomedical contexts.

Our investigations demonstrated that the effectiveness of BiomedCLIP fine-tuning is highly sensitive to the chosen methodology. We found that minimal fine-tuning (e.g., around 5 epochs) using masked images and concise textual prompts often yielded robust saliency maps. This effectiveness likely stems from several factors: using masked images focuses the model on relevant features; shorter training durations prevent overfitting, promoting generalization; and concise prompts provide a clearer, less ambiguous signal to the model.

The refined saliency maps generated by the optimally tuned BiomedCLIP serve as a more effective foundation for deriving precise spatial prompts (points or bounding boxes) intended for use with SAM. The experiments underscored the importance of the prompt selection method, showing that simplistic strategies (like selecting random points from only the largest contour) are not universally optimal, especially for sparse or distributed ROIs where multi-contour strategies are preferable. The effectiveness of any prompt selection strategy is intrinsically linked to the quality and characteristics of the saliency map produced by the fine-tuned vision-language model.

Furthermore, while this work focused primarily on optimizing and understanding the reproducibility of the prompt generation stage via BiomedCLIP, achieving reliable end-to-end segmentation necessitates acknowledging factors beyond this scope. Potential stochasticity within SAM’s internal mask decoding process or implementation variations could still affect final segmentation reproducibility even with identical input prompts, representing an area for future investigation relevant to the broader goal of creating robust biomedical tools. Our results, focused on improving the input to SAM, indirectly support the notion that an off-the-shelf SAM could be utilized more effectively when provided with high-quality, targeted prompts derived from a well-tuned upstream model, particularly for tasks where ROIs are clearly indicated by the prompts. However, the question of whether SAM itself requires fine-tuning for optimal performance across diverse biomedical conditions remains an open area addressed by other studies in the field.

In conclusion, a carefully fine-tuned vision-language model, configured with optimized parameters and training strategies, coupled with intelligent, context-aware selection of spatial prompts derived from its output, enhances the potential for effective segmentation using an off-the-shelf SAM by providing it with improved guidance. Future work stemming from this investigation should focus on the prompt generation and selection process:

- Developing more sophisticated, adaptive, and potentially automated spatial prompt selection algorithms that optimally leverage the information in the refined saliency maps, tailored to specific biomedical imaging characteristics and ROI distributions.

- Exploring advanced fine-tuning techniques (e.g., parameter-efficient methods) specifically for the vision-language model (BiomedCLIP) to further optimize saliency map quality and stability.
- Investigating how variations in the quality and type of prompts generated by different BiomedCLIP configurations influence the downstream performance and behavior of a fixed SAM model.

By improving the crucial link between textual/visual understanding (BiomedCLIP) and spatial segmentation (SAM), such advancements will contribute to realizing more robust, reliable, and clinically valuable biomedical image segmentation frameworks that leverage the power of foundation models.

CHAPTER 6

CONCLUSION

6.1 Summary of Contributions

This dissertation systematically addressed key challenges in biomedical image segmentation by developing and evaluating supervised, self-supervised, minimally supervised, and foundation model-based segmentation methods. Collectively, these chapters form a cohesive framework intended to advance segmentation accuracy, generalizability, and usability across diverse biomedical imaging scenarios.

In Chapter 2, we introduced TSeg, a comprehensive pipeline for 3D cell segmentation, tracking, and motility analysis, demonstrated with *Toxoplasma gondii*. TSeg integrates established CNN-based segmentation tools (CellPose, PlantSeg) with tracking and analysis modules within a user-friendly Napari plugin, enabling researchers with limited coding expertise to deploy these methods. Its performance evaluation on Cell Tracking Challenge (CTC) datasets highlights its potential as an applicable tool in biomedical research requiring 3D analysis.

Chapter 3 presented a self-supervised segmentation method for cilia, generating pseudo-labels via optical flow and autoregressive modeling of motion patterns. This approach reduces the reliance on manually annotated datasets for training segmentation models while achieving reasonable performance, particularly in identifying ciliary regions (comparable sensitivity to supervised methods). Although demonstrated specifically for cilia segmentation, the core idea of using motion signatures for pseudo-labeling could potentially be extended to other time-series biomedical imaging tasks, such as cardiac motion analysis or dynamic cell processes.

In Chapter 4, we leveraged contrastive learning for minimally supervised segmentation, requiring only simple point-based user interaction to define regions of interest. This self-supervision approach learns discriminative representations from unlabeled temporal data, enabling segmentation with minimal annotation effort. The framework showed efficacy across various 2D CTC datasets, illustrating the potential of contrastive methods for tackling data scarcity. One specific future direction is to investigate whether the contrastive coding strategy developed here could serve as an alternative to standard contrastive loss functions (e.g., DHN-NCE loss) in vision-language models like BiomedCLIP, potentially improving their fine-tuning for biomedical applications.

Chapter 5 explored the role of foundation models, specifically focusing on fine-tuning the BiomedCLIP vision-language model to improve the generation of spatial prompts for Segment Anything Models (SAMs). We investigated the impact of fine-tuning strategies, data masking, prompt consistency, and hyperparameters on the quality of generated saliency maps. While SAMs provide flexible segmentation solutions, our findings regarding the upstream prompt generation highlight that domain-specific adaptation, even if minimal, is often beneficial for reliable performance on specialized tasks like cilia identification. Nonetheless, the accessibility of foundation models like SAM, potentially guided by outputs from models like BiomedCLIP, opens avenues for broader adoption of advanced segmentation tools in biomedical image analysis.

6.2 Theoretical and Practical Implications

The proposed methodologies contribute both theoretically and practically to biomedical image segmentation. TSeg serves as an integrated segmentation and tracking solution applicable to various 3D cellular imaging tasks. Its GUI-driven design lowers the barrier to entry for non-experts, potentially accelerating the adoption of deep learning techniques in biological research.

The self-supervised motion-based segmentation method (Chapter 3) introduces an efficient way of generating training data by leveraging inherent temporal dynamics, reducing annotation workload without eliminating the need for model training itself. This makes it well-suited for time-series imaging tasks where motion is a key characteristic.

Our contrastive learning framework (Chapter 4) further illustrates the efficacy of self-supervision, learning useful representations directly from unlabeled image sequences with minimal interactive guidance. Exploring its integration into large-scale biomedical models could potentially enhance their ability to capture subtle morphological differences.

Finally, the investigation of foundation models (Chapter 5) underlines their dual nature: promise coupled with limitations in specialized domains. While generic pretraining might overlook domain-specific nuances, appropriate adaptation (here, of the prompt-generating model) can yield competitive results. The ability of foundation models like SAM to deliver usable segmentations with minimal direct supervision or fine-tuning (when provided good prompts) underscores their potential for widespread application.

6.3 Limitations, Failure Modes, and Future Directions

Despite the advancements presented, several challenges and limitations warrant discussion. Acknowledging these is crucial for responsible application and identifying pathways for future improvement, while not negating the utility of the methods in appropriate contexts.

Limitations and Failure Modes

Data Dependency and Domain Shift: The performance of all developed methods remains partly dependent on the characteristics of the training data. Domain shifts (e.g., applying a model trained on one microscope to data from another) or datasets with high heterogeneity can degrade accuracy. This is a common challenge in machine learning but particularly pronounced in biomedical imaging due to variability in equipment, protocols, and biological samples.

Method-Specific Failures:

- *TSeg (Chapter 2):* Performance is constrained by the chosen backend (CellPose/PlantSeg) and computational resources, especially for large 3D+time datasets. The tracking module struggles with complex cellular events like dense overlaps, division, or fusion, potentially leading to trajectory errors.
- *Self-Supervised Cilia Segmentation (Chapter 3):* Accuracy relies on the quality of motion-derived pseudo-labels. Extraneous motion (e.g., stage drift) can create false signals, while very slow or irregular cilia movement may not generate strong enough signals, leading to under-segmentation. Training solely on motile cilia might also bias the model if dyskinetic cilia have distinct static appearances not captured during training.
- *Contrastive Learning Segmentation (Chapter 4):* This method struggled with images containing artifacts visually similar to the target texture (e.g., hydrogel shadows in BF-C2DL datasets) leading to false positives. Performance also decreased in low-contrast images (Fluo-N2DL-HeLa) or where cell size and internal texture variation were large compared to the patch size (DIC-C2DH-HeLa), leading to incomplete masks or imprecise boundaries.
- *Foundation Model Prompting (BiomedCLIP, Chapter 5):* While fine-tuning improved saliency maps, it increased false positives. The system remains sensitive to prompt phrasing and hyperparameter choices, impacting reproducibility of the prompt generation stage. The quality of the final SAM segmentation (not performed in Chapter 5) would depend heavily on these upstream factors.

Need for Validation: While self-supervision reduces manual annotation *for training*, expert validation of the final segmentation outputs remains crucial for biological and clinical relevance, especially in diagnostic settings.

Scalability: Processing large 3D or 4D datasets remains computationally intensive for methods like TSeg. Real-time application may require further optimization or specialized hardware.

Uniqueness of Failures in Biomedical Images

Many failures observed (e.g., sensitivity to low contrast, texture variations, similar artifacts) are exacerbated in biomedical imaging compared to general computer vision. Reasons include the inherent complexity and subtlety of biological structures, lack of sharp canonical boundaries for many cell types or tissues, prevalence of imaging artifacts, and inter-sample biological variability.

Future Directions

Building upon this work, future research could pursue several avenues:

Improving Scalability and Robustness:

- *TSeg Scalability*: Investigate computational optimizations like data tiling, sparsification techniques, asynchronous processing, or leveraging cloud/HPC resources. Explore alternative, potentially lighter-weight segmentation backends.
- *Contrastive Learning Boundaries*: Enhance boundary definition by incorporating multi-scale patch analysis, exploring boundary-specific loss terms, refining negative sampling strategies, or integrating minimal boundary annotations.

Exploring Advanced Architectures: While TSeg currently uses CNNs (U-Nets via CellPose/PlantSeg), future iterations *could* explore transformer-based architectures (like Swin-UNET, UNETR) which have shown promise in capturing long-range spatial dependencies in medical images, potentially improving segmentation of complex or large structures. However, this work did not implement or evaluate transformers within TSeg.

Enhancing Generalization and Applicability:

- *Additional Datasets*: Validate methods on a wider range of datasets beyond CTC or the specific cilia data, such as imaging data from different organisms, other microscopy modalities (e.g., Electron Microscopy, Confocal), different disease states, or clinical imaging archives (e.g., radiology scans for foundation models).
- *Multimodal Data Integration*: Explore fusion techniques to combine imaging data with other sources. For example, integrating clinical metadata or molecular profiling could potentially guide segmentation models via attention mechanisms or be used as conditional inputs to improve specificity.

Leveraging Newer AI Paradigms:

- *Foundation Models*: Investigate newer foundation models possessing enhanced reasoning capabilities or explicit knowledge integration, potentially improving segmentation accuracy through better contextual understanding.
- *Reinforcement Learning*: Explore RL for optimizing interactive segmentation workflows (e.g., learning the best sequence of user prompts/corrections) or for automated hyperparameter tuning.

Improving Trust and Interpretability: Incorporate uncertainty estimation techniques (e.g., Bayesian deep learning, ensemble methods) to provide confidence scores for segmentations, and utilize explainability methods (e.g., attention maps, Grad-CAM) to understand model decision-making, enhancing clinical acceptance.

6.4 Broader Impact and Scientific Contributions

This dissertation advances biomedical image segmentation by introducing tools and methods intended to improve accessibility, reduce annotation demands, and enhance model robustness. The potential impact spans multiple domains:

- **Clinical Diagnostics:** Automated segmentation can facilitate the detection and analysis of conditions such as ciliopathies, cancer, and neurodegenerative disorders, potentially streamlining clinical workflows and improving patient care.
- **Biomedical Research:** The proposed segmentation pipelines can be applied to study cellular morphology, motility patterns, and disease progression, potentially accelerating research in areas like developmental biology, immunology, and infectious disease.
- **AI in Healthcare:** By making segmentation tools more accessible and less reliant on large annotated datasets, this work supports broader adoption of AI-driven diagnostics and personalized medicine, potentially transforming approaches to healthcare challenges.

In conclusion, these contributions establish a foundation for developing more scalable, efficient, and interpretable biomedical segmentation approaches. By integrating deep learning, self-supervision, and foundation models, this dissertation aims to bring the field closer to practical, domain-ready, and data-efficient solutions that can serve a wide range of medical and biological research endeavors.

APPENDIX A

APPENDICES

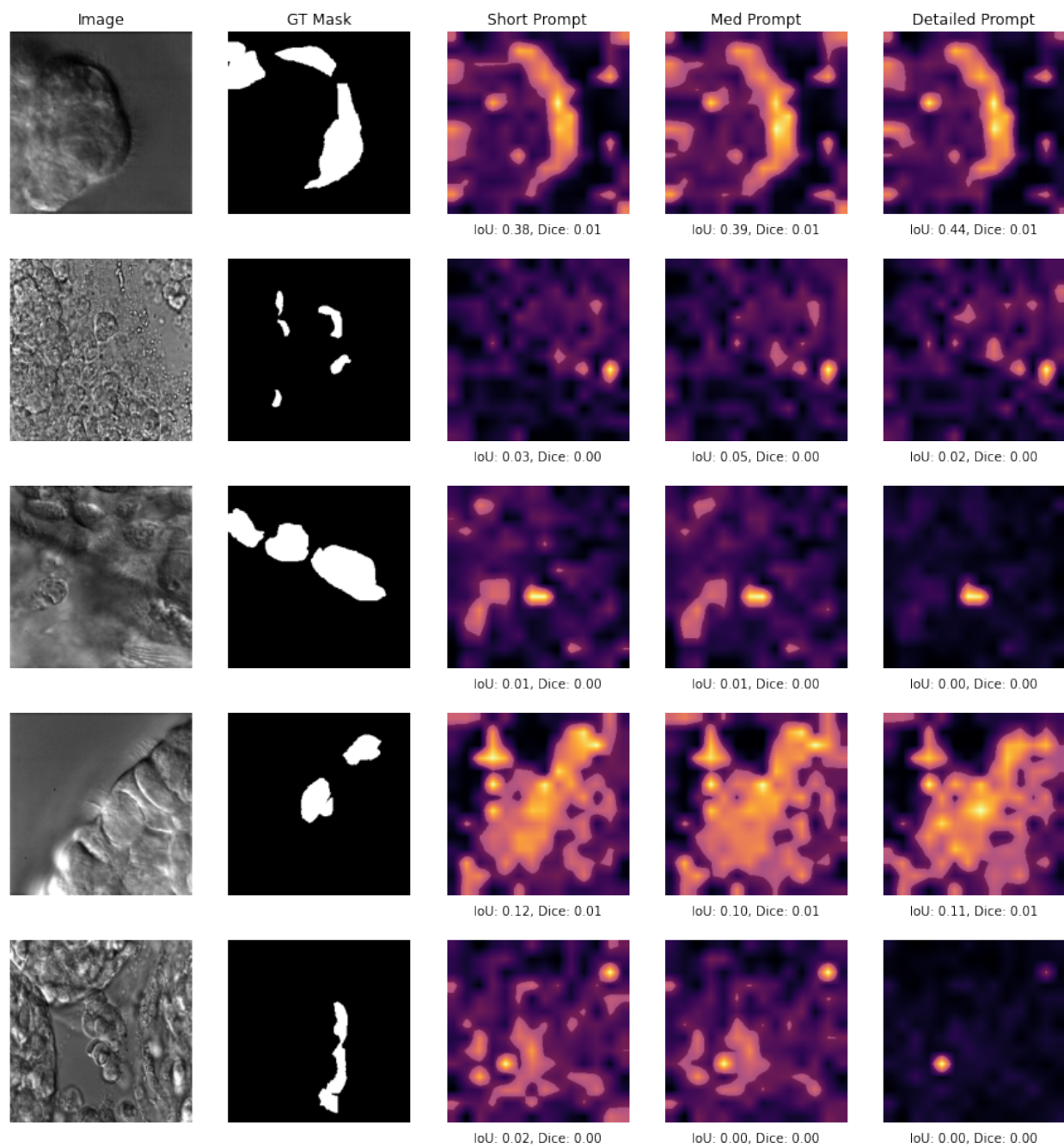


Figure A.1: Improved ciliary region localization after fine-tuning BiomedCLIP for one epoch using masked images with short captions.

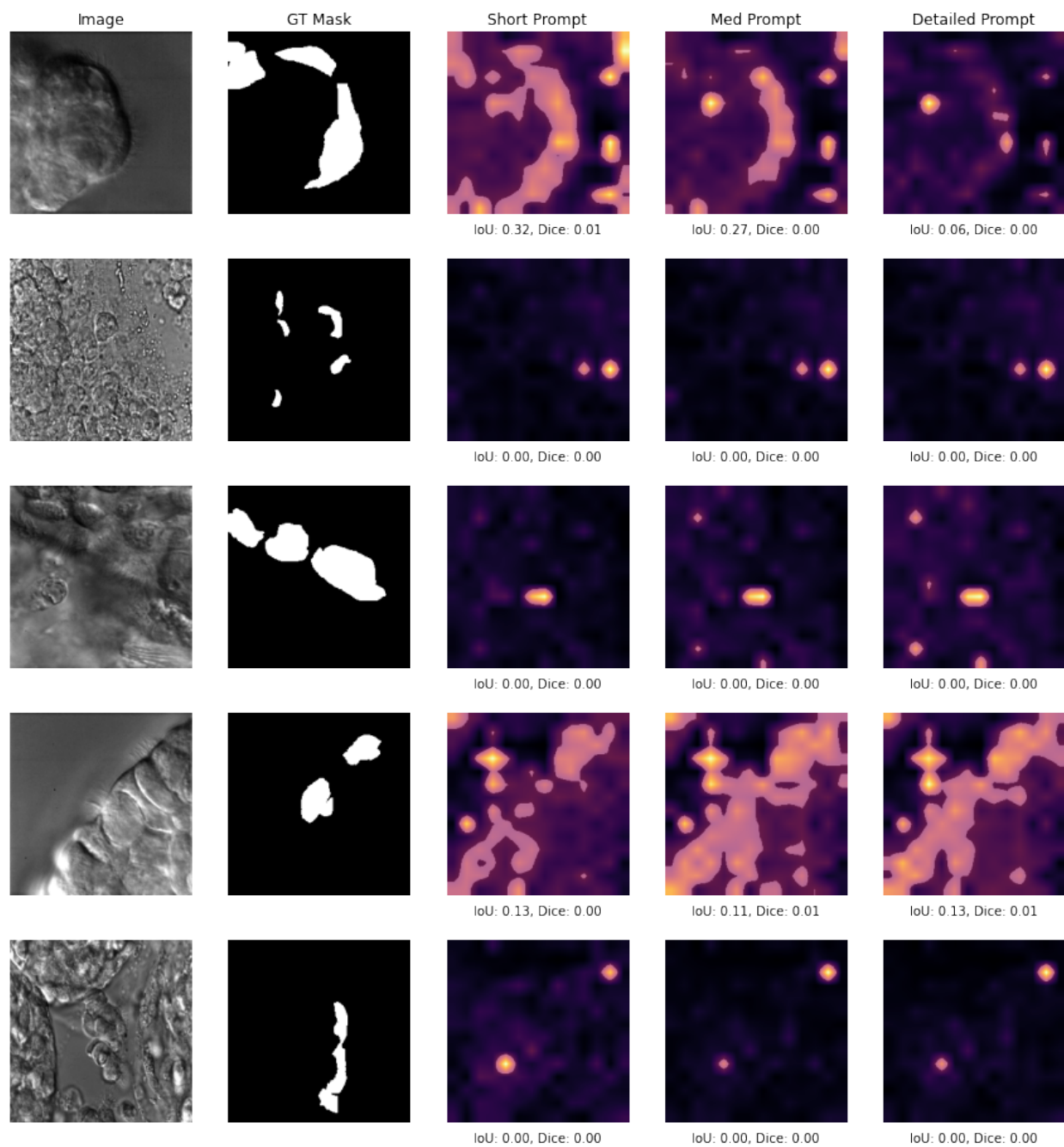


Figure A.2: Effective and balanced localization achieved by fine-tuning BiomedCLIP with a reduced dataset size (small train) using masked images and short captions for 5 epochs.

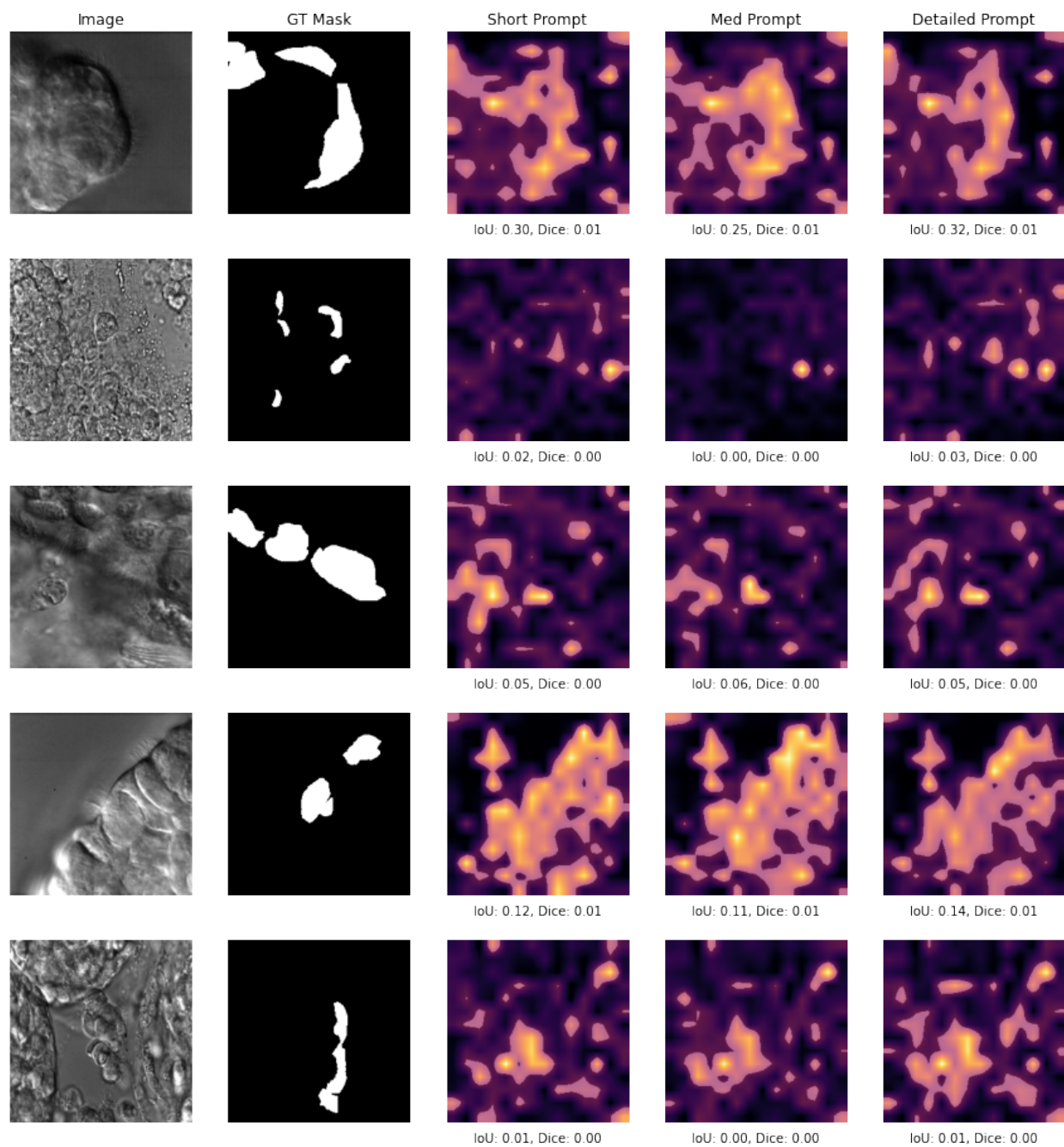


Figure A.3: Localization results demonstrating increased flexibility but higher false positives when fine-tuning BiomedCLIP using long captions.

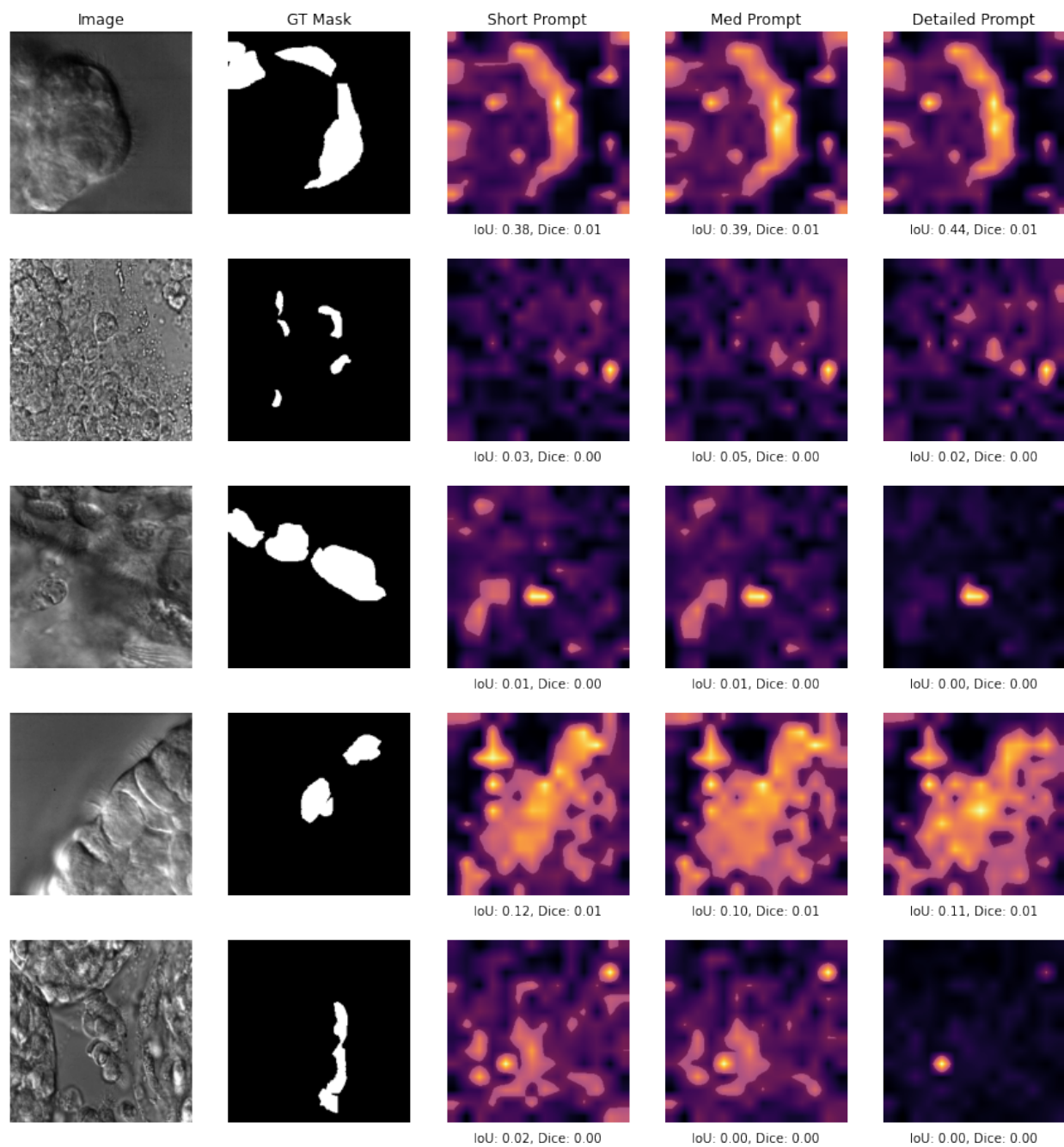


Figure A.4: Localization impact of higher temperature ($\tau = 0.4$) during fine-tuning, resulting in blurred similarity distributions and compromised accuracy.

Table A.1: Cell Tracking Challenge Datasets

Dataset Name	Organism	Description	Imaging Modality	Dimension
BF-C2DL-HSC	Mouse	Hematopoietic stem cells cultured in hydrogel microwells.	Brightfield Microscopy	2D
BF-C2DL-MuSC	Mouse	Muscle stem cells cultured in hydrogel microwells.	Brightfield Microscopy	2D
DIC-C2DH-HeLa	Human	HeLa cells cultured on a flat glass surface.	Differential Interference Contrast (DIC) Microscopy	2D
Fluo-C2DL-Huh7	Human	Huh7 cells expressing the fusion protein YFP-TIA-1.	Fluorescence Microscopy	2D
Fluo-C2DL-MSC	Rat	Mesenchymal stem cells cultured on a flat polyacrylamide substrate.	Fluorescence Microscopy	2D
Fluo-N2DH-GOWT1	Mouse	GFP-GOWT1 stem cells.	Fluorescence Microscopy	2D
Fluo-N2DL-HeLa	Human	HeLa cells stably expressing H2b-GFP.	Fluorescence Microscopy	2D
Fluo-C3DH-A549	Human	A549 lung cancer cells embedded in a Matrigel matrix.	Fluorescence Microscopy	3D
Fluo-C3DH-H157	Human	GFP-transfected H157 lung cancer cells embedded in a Matrigel matrix.	Fluorescence Microscopy	3D
Fluo-C3DL-MDA231	Human	MDA231 human breast carcinoma cells infected with a pMSCV vector including the GFP sequence, embedded in a collagen matrix.	Fluorescence Microscopy	3D
Fluo-N3DH-CE	C. elegans	Developing C. elegans embryo.	Fluorescence Microscopy	3D
Fluo-N3DH-CHO	Chinese Hamster	Chinese Hamster Ovarian (CHO) nuclei overexpressing GFP-PCNA.	Fluorescence Microscopy	3D
Fluo-N3DL-DRO	Drosophila melanogaster	Developing Drosophila melanogaster embryo.	Fluorescence Microscopy	3D
Fluo-N3DL-TRIC	Tribolium castaneum	Developing Tribolium castaneum embryo (3D cartographic projection).	Fluorescence Microscopy	3D
Fluo-N3DL-TRIF	Tribolium castaneum	Developing Tribolium castaneum embryo.	Fluorescence Microscopy	3D

Table A.2: Overview of Cell Tracking Challenge Datasets with Sample Images.

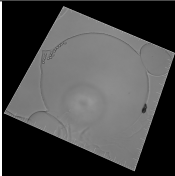
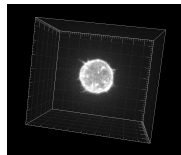
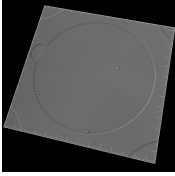
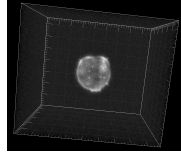
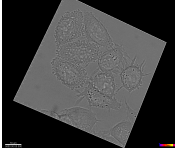
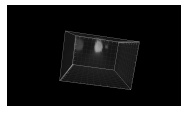
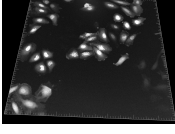
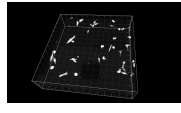
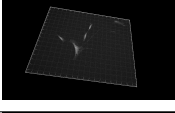
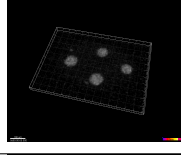
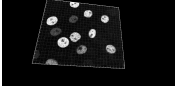
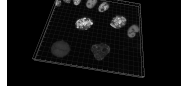
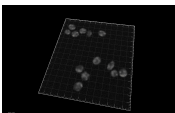
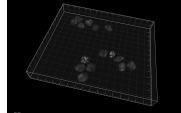
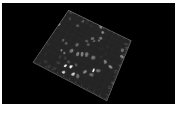
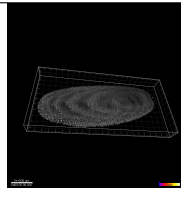
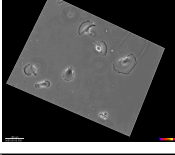
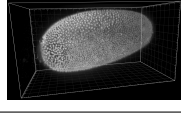
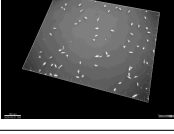
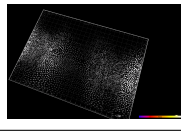
2D Dataset Name	Sample Image	3D Dataset Name	Sample Image
BF-C2DL-HSC		Fluo-C3DH-A549	
BF-C2DL-MuSC		Fluo-C3DH-A549-SIM	
DIC-C2DH-HeLa		Fluo-C3DH-H157	
Fluo-C2DL-Huh7		Fluo-C3DL-MDA231	
Fluo-C2DL-MSC		Fluo-N3DH-CE	
Fluo-N2DH-GOWT1		Fluo-N3DH-CHO	
Fluo-N2DH-SIM		Fluo-N3DH-SIM	
Fluo-N2DL-HeLa		Fluo-N3DL-DRO	
PhC-C2DH-U373		Fluo-N3DL-TRIF	
PhC-C2DL-PSC		Fluo-N3DL-TRIC	

Table A.3: CellPose Performance on 2D Datasets

dataset name	BF-C2DL-HSC	BF-C2DL-MuSC	DIC-C2DH-HeLa	Fluo-C2DL-Huh7	Fluo-C2DL-MSC	Fluo-N2DH-GOWTi	Fluo-N2DH-SIM+	Fluo-N2DL-HeLa	PhC-C2DH-U373	PhC-C2DL-PSC
cyto3	0.98	0.95	0.36	0.60	0.90	0.89	0.82	0.75	0.87	0.87
nuclei	0.99	0.99	0.36	0.60	0.89	0.86	0.80	0.75	0.87	0.90
cyto2_cp3	0.94	0.95	0.35	0.60	0.89	0.87	0.82	0.74	0.85	0.87
tissuenet_cp3	0.98	0.97	0.36	0.59	0.89	0.86	0.80	0.75	0.87	0.91
livecell_cp3	0.98	0.99	0.36	0.59	0.89	0.86	0.80	0.75	0.87	0.91
yeast_PhC_cp3	0.98	0.98	0.34	0.59	0.84	0.86	0.71	0.75	0.86	0.88
yeast_BF_cp3	0.99	0.99	0.36	0.60	0.89	0.86	0.79	0.73	0.87	0.91
bact_phase_cp3	0.97	0.97	0.35	0.59	0.89	0.86	0.79	0.73	0.86	0.89
bact_fluor_cp3	0.93	0.95	0.33	0.59	0.89	0.86	0.79	0.73	0.84	0.89
deepbacs_cp3	0.99	0.99	0.36	0.60	0.89	0.86	0.80	0.75	0.87	0.90
cyto2	0.95	0.96	0.36	0.60	0.90	0.88	0.82	0.73	0.86	0.86
cyto	0.97	0.97	0.35	0.59	0.89	0.87	0.80	0.72	0.84	0.86
CPx	0.97	0.98	0.34	0.59	0.89	0.88	0.80	0.73	0.86	0.90
neurips_grayscale_cyto2	0.99	0.98	0.35	0.60	0.89	0.88	0.89	0.67	0.87	0.91
CP	0.97	0.95	0.35	0.60	0.89	0.88	0.81	0.72	0.86	0.89
TN1	0.98	0.99	0.36	0.60	0.89	0.86	0.80	0.75	0.86	0.91
TN2	0.99	0.98	0.36	0.59	0.89	0.86	0.80	0.75	0.87	0.91
TN3	0.99	0.99	0.36	0.60	0.89	0.86	0.80	0.75	0.87	0.91
LC1	0.99	0.99	0.27	0.59	0.89	0.89	0.79	0.75	0.87	0.91
LC2	0.96	0.97	0.36	0.60	0.89	0.86	0.79	0.75	0.85	0.90
LC3	0.98	0.97	0.33	0.60	0.89	0.86	0.80	0.74	0.82	0.85
LC4	0.95	0.98	0.36	0.59	0.89	0.86	0.80	0.75	0.87	0.83

Table A.4: PlantSeg Performance on 2D Datasets

dataset name	BF-C2DL-HSC	BF-C2DL-MuSC	DIC-C2DH-HeLa	Fluo-C2DL-Huh7	Fluo-C2DL-MSK	Fluo-N2DH-GOWT1	Fluo-N2DH-SIM+	Fluo-N2DL-HeLa	PhC-C2DH-U373	PhC-C2DL-PSC
confocal_2D_unet_ovules_ds2x	0.99	0.99	0.36	0.59	0.89	0.86	0.80	0.75	0.87	0.91
lightsheet_2D_unet_root_ds1x	0.99	0.99	0.36	0.59	0.89	0.86	0.80	0.75	0.87	0.91
lightsheet_2D_unet_root_nuclei_ds1x	0.99	0.99	0.36	0.59	0.89	0.86	0.80	0.75	0.87	0.91
confocal_2D_unet_sa_meristem_cells	0.99	0.99	0.36	0.59	0.89	0.86	0.80	0.75	0.87	0.91

Table A.5: CellPose Performance on 3D Datasets

dataset_name	Fluo-C3DH-A549	Fluo-C3DH-A549-SIM	Fluo-C3DH-H157	Fluo-N3DH-CE	Fluo-N3DH-CHO	Fluo-N3DH-SIM+
cyto3	0.96	0.97	0.93	0.80	0.83	0.92
nuclei	0.96	0.97	0.93	0.80	0.84	0.93
cyto2_cp3	0.97	0.97	0.93	0.80	0.79	0.93
tissuenet_cp3	0.96	0.97	0.93	0.80	0.84	0.93
livecell_cp3	0.96	0.95	0.93	0.80	0.84	0.93
yeast_PhC_cp3	0.92	0.93	0.88	0.68	0.84	0.86
yeast_BF_cp3	0.95	0.97	0.92	0.80	0.83	0.88
bact_phase_cp3	0.96	0.97	0.93	0.79	0.83	0.93
bact_fluor_cp3	0.96	0.97	0.93	0.80	0.83	0.93
deepbacs_cp3	0.96	0.97	0.93	0.80	0.84	0.93
cyto2	0.96	0.98	0.93	0.80	0.84	0.94
cyto	0.96	0.97	0.93	0.80	0.82	0.93
CPx	0.98	0.99	0.93	0.80	0.83	0.93
neurips_grayscale_cyto2	0.97	0.98	0.93	0.73	0.82	0.92
CP	0.98	0.99	0.93	0.79	0.82	0.93
TN1	0.96	0.97	0.93	0.80	0.84	0.93
TN2	0.96	0.97	0.93	0.80	0.84	0.93
TN3	0.96	0.99	0.93	0.80	0.84	0.93
LC1	0.86	0.97	0.92	0.80	0.82	0.92
LC2	0.96	0.97	0.93	0.80	0.84	0.93
LC3	0.96	0.13	0.93	0.79	0.82	0.30
LC4	0.96	0.97	0.93	0.80	0.84	0.93

Table A.6: PlantSeg Performance on 3D Datasets

dataset_name	Fluo-C₃DH-A549	Fluo-C₃DH-A549-SIM	Fluo-C₃DH-H157	Fluo-N₃DH-CE	Fluo-N₃DH-CHO	Fluo-N₃DH-SIM+
generic_confocal_3D_unet	0.96	0.97	0.88	0.78	0.84	—
generic_light_sheet_3D_unet	0.96	0.97	0.88	0.78	0.84	—
confocal_3D_unet_ovules_ds1x	0.96	0.97	0.88	0.78	0.84	—
confocal_3D_unet_ovules_ds2x	0.96	0.97	0.88	0.78	0.84	—
confocal_3D_unet_ovules_ds3x	0.96	0.97	0.88	—	0.84	—
lightsheet_3D_unet_root_ds1x	0.96	0.97	0.88	—	0.84	—
lightsheet_3D_unet_root_ds2x	0.96	0.97	0.88	—	0.84	—
lightsheet_3D_unet_root_ds3x	0.96	0.97	0.88	—	0.84	—
lightsheet_3D_unet_root_nuclei_ds1x	0.96	0.97	0.88	—	0.84	—
confocal_3D_unet_sa_meristem_cells	0.96	0.97	0.88	—	0.84	—
confocal_3D_unet_mouse_embryo_nuclei	0.96	0.97	0.88	—	0.84	—
PlantSeg_3Dnuc_platinum	0.96	0.97	0.88	—	0.84	—

BIBLIOGRAPHY

- [1] M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, “Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation,” *arXiv preprint arXiv:1802.06955*, 2018.
- [2] R. Biswas, “Polyp-sam++: Can a text guided sam perform better for polyp segmentation?” *arXiv preprint arXiv:2308.06623*, 2023.
- [3] D. Bolya, C. Ryali, J. Hoffman, and C. Feichtenhofer, “Window attention is bugged: How not to interpolate position embeddings,” *arXiv preprint arXiv:2311.05613*, 2023.
- [4] L. Breiman, “Random forests,” *Machine learning*, vol. 45, pp. 5–32, 2001.
- [5] Z. Cai and N. Vasconcelos, “Cascade r-cnn: Delving into high quality object detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6154–6162.
- [6] H. Cao et al., “Swin-unet: Unet-like pure transformer for medical image segmentation,” in *European conference on computer vision*, Springer, 2022, pp. 205–218.
- [7] L. von Chamier et al., “Democratising deep learning for microscopy with zerocostdl4mic,” *Nature communications*, vol. 12, no. 1, pp. 1–18, 2021. DOI: 10.1038/s41467-021-22518-0.
- [8] C. Chen et al., “Improving the generalizability of convolutional neural network-based segmentation on cmr images,” *Frontiers in Cardiovascular Medicine*, vol. 7, 2020, ISSN: 2297-055X. DOI: <https://doi.org/10.3389/fcvm.2020.00105>. [Online]. Available: <https://www.frontiersin.org/articles/10.3389/fcvm.2020.00105>.
- [9] J. Chen et al., “Transunet: Transformers make strong encoders for medical image segmentation,” *arXiv preprint arXiv:2102.04306*, 2021.
- [10] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *International conference on machine learning*, PmLR, 2020, pp. 1597–1607.
- [11] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, “Big self-supervised models are strong semi-supervised learners,” *Advances in neural information processing systems*, vol. 33, pp. 22 243–22 255, 2020.
- [12] X. Chen, H. Fan, R. Girshick, and K. He, “Improved baselines with momentum contrastive learning,” *arXiv preprint arXiv:2003.04297*, 2020.

- [13] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, “3d u-net: Learning dense volumetric segmentation from sparse annotation,” in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17–21, 2016, Proceedings, Part II 19*, Springer, 2016, pp. 424–432.
- [14] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine learning*, vol. 20, pp. 273–297, 1995.
- [15] I. Croitoru, S.-V. Bogolin, and M. Leordeanu, “Unsupervised learning of foreground object segmentation,” *International Journal of Computer Vision*, vol. 127, pp. 1279–1302, 2019.
- [16] G. Doretto, A. Chiuso, Y. N. Wu, and S. Soatto, “Dynamic textures,” *International journal of computer vision*, vol. 51, pp. 91–109, 2003. DOI: <https://doi.org/10.1023/A:1021669406132>.
- [17] P. de Dumast and M. B. Cuadra, “Domain generalization in fetal brain mri segmentation with multi-reconstruction augmentation,” in *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, IEEE, 2023, pp. 1–5.
- [18] E. Fazeli et al., “Automated cell tracking using stardist and trackmate,” *F1000Research*, vol. 9, 2020. DOI: [10.12688/f1000research.27019.1](https://doi.org/10.12688/f1000research.27019.1).
- [19] M. S. Fazli, S. A. Velia, S. N. Moreno, and S. Quinn, “Unsupervised discovery of toxoplasma gondii motility phenotypes,” in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, IEEE, 2018, pp. 981–984. DOI: [10.1109/isbi.2018.8363735](https://doi.org/10.1109/isbi.2018.8363735).
- [20] M. S. Fazli, S. A. Vella, S. N. Moreno, G. E. Ward, and S. P. Quinn, “Toward simple & scalable 3d cell tracking,” in *2018 IEEE International Conference on Big Data (Big Data)*, IEEE, 2018, pp. 3217–3225. DOI: [10.1109/BigData.2018.8622403](https://doi.org/10.1109/BigData.2018.8622403).
- [21] M. S. Fazli et al., “Lightweight and scalable particle tracking and motion clustering of 3d cell trajectories,” in *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, IEEE, 2019, pp. 412–421. DOI: [10.1109/dsaa.2019.00056](https://doi.org/10.1109/dsaa.2019.00056).
- [22] I. Goodfellow et al., “Generative adversarial nets,” *Advances in neural information processing systems*, vol. 27, 2014. DOI: [10.48550/arXiv.1406.2661](https://doi.org/10.48550/arXiv.1406.2661).
- [23] Y. Gu et al., “Domain-specific language model pretraining for biomedical natural language processing,” *ACM Transactions on Computing for Healthcare (HEALTH)*, vol. 3, no. 1, pp. 1–23, 2021.
- [24] S. Guan, A. A. Khan, S. Sikdar, and P. V. Chitnis, “Fully dense unet for 2-d sparse photoacoustic tomography artifact removal,” *IEEE journal of biomedical and health informatics*, vol. 24, no. 2, pp. 568–576, 2019.
- [25] J. N. Hansen, S. Rassmann, B. Stüven, N. Jurisch-Yaksi, and D. Wachten, “CiliaQ: A simple, open-source software for automated quantification of ciliary morphology and fluorescence in 2d, 3d, and 4D images,” *The European Physical Journal E*, vol. 44, no. 2, p. 18, Mar. 2021. DOI: <https://doi.org/10.1140/epje/s10189-021-00031-y>.

- [26] A. Hatamizadeh et al., “Unetr: Transformers for 3d medical image segmentation,” in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 574–584.
- [27] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, “Masked autoencoders are scalable vision learners,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.
- [28] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.
- [29] K. He, G. Gkioxari, P. Dollár, and R. Girshick, *Mask r-cnn*, 2018. arXiv: 1703.06870 [cs.CV]. [Online]. Available: <https://arxiv.org/abs/1703.06870>.
- [30] S. Hoyer-Fender, “Primary and motile cilia: Their ultrastructure and ciliogenesis,” in *Cilia and Nervous System Development and Function*, K. L. Tucker and T. Caspary, Eds. Dordrecht: Springer Netherlands, 2013, pp. 1–53, ISBN: 978-94-007-5808-7. DOI: 10.1007/978-94-007-5808-7_1. [Online]. Available: https://doi.org/10.1007/978-94-007-5808-7_1.
- [31] E. J. Hu et al., “Lora: Low-rank adaptation of large language models,” *ICLR*, vol. 1, no. 2, p. 3, 2022.
- [32] J. Huang et al., “Learning to prompt segment anything models,” *arXiv preprint arXiv:2401.04651*, 2024.
- [33] X. Huang et al., “Segment and caption anything,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 13 405–13 417.
- [34] Z. Huang, L. Lin, P. Cheng, L. Peng, and X. Tang, “Multi-modal brain tumor segmentation via missing modality synthesis and modality-level attention fusion,” *arXiv preprint arXiv:2203.04586*, 2022.
- [35] M. L. Hutchinson, E. Antono, B. M. Gibbons, S. Paradiso, J. Ling, and B. Meredig, *Overcoming data scarcity with transfer learning*, 2017. DOI: <https://doi.org/10.48550/arXiv.1711.05099>. arXiv: 1711.05099 [cs.LG].
- [36] M. Hyndman, A. D. Jepson, and D. J. Fleet, “Higher-order autoregressive models for dynamic textures,” in *BMVC*, 2007, pp. 1–10.
- [37] P. Iakubovskii, *Segmentation models pytorch*, https://github.com/qubvel/segmentation_models.pytorch, 2019.
- [38] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, “Nnu-net: A self-configuring method for deep learning-based biomedical image segmentation,” *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.
- [39] F. Isensee et al., “Nnu-net revisited: A call for rigorous validation in 3d medical image segmentation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2024, pp. 488–498.

- [40] X. Ji, J. F. Henriques, and A. Vedaldi, “Invariant information clustering for unsupervised image classification and segmentation,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9865–9874.
- [41] V. Kapoor and C. Carabaña, “Cell tracking in 3d using deep learning segmentations,” in *Python in Science Conference*, 2021, pp. 154–161. DOI: 10.25080/majora-1b6fd038-014.
- [42] A. Kar, M. Petit, Y. Refahi, G. Cerutti, C. Godin, and J. Traas, “Assessment of deep learning algorithms for 3d instance segmentation of confocal image datasets,” *bioRxiv*, 2021. DOI: 10.1101/2021.06.09.447748. eprint: <https://www.biorxiv.org/content/early/2021/06/10/2021.06.09.447748.full.pdf>. [Online]. Available: <https://www.biorxiv.org/content/early/2021/06/10/2021.06.09.447748>.
- [43] C. Kempeneers and M. A. Chilvers, “To beat, or not to beat, that is question! the spectrum of ciliopathies,” *Pediatric Pulmonology*, vol. 53, no. 8, pp. 1122–1129, 2018. DOI: <https://doi.org/10.1002/ppul.24078>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/ppul.24078>. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/ppul.24078>.
- [44] T. Khatibi, N. Rezaei, L. Ataei Fashtami, and M. Totonchi, “Proposing a novel unsupervised stack ensemble of deep and conventional image segmentation (sedcis) method for localizing vitiligo lesions in skin images,” *Skin Research and Technology*, vol. 27, no. 2, pp. 126–137, 2021. DOI: <http://dx.doi.org/10.1111/srt.12920>.
- [45] D. Kim, D. Cho, and I. S. Kweon, “Self-supervised video representation learning with space-time cubic puzzles,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, 2019, pp. 8545–8552. DOI: <https://doi.org/10.48550/arXiv.1811.09795>.
- [46] S. Kim et al., *Medivista: Medical video segmentation via temporal fusion sam adaptation for echocardiography*, 2024. arXiv: 2309.13539 [eess.IV]. [Online]. Available: <https://arxiv.org/abs/2309.13539>.
- [47] S. Kim et al., “Medivista: Medical video segmentation via temporal fusion sam adaptation for echocardiography,” *IEEE Journal of Biomedical and Health Informatics*, 2025.
- [48] A. Kirillov, K. He, R. Girshick, and P. Dollár, “A unified architecture for instance and semantic segmentation,” in *Computer Vision and Pattern Recognition Conference*, CVPR, 2017. DOI: <https://doi.org/10.48550/arXiv.2112.04603>.
- [49] A. Kirillov et al., “Segment anything,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015–4026. arXiv: 2304.02643 [cs.CV]. [Online]. Available: <https://arxiv.org/abs/2304.02643>.
- [50] T. Koleilat, H. Asgariandehkordi, H. Rivaz, and Y. Xiao, “Medclip-sam: Bridging text and image towards universal medical image segmentation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2024, pp. 643–653.

- [51] T. Koleilat, H. Asgariandehkordi, H. Rivaz, and Y. Xiao, "Medclip-samv2: Towards universal text-driven medical image segmentation," *arXiv preprint arXiv:2409.19483*, 2024.
- [52] A. Kolesnikov, X. Zhai, and L. Beyer, "Revisiting self-supervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 1920–1929. DOI: <https://doi.org/10.48550/arXiv.1901.09005>.
- [53] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012. DOI: 10.1145/3065386.
- [54] J. Krois et al., "Generalizability of deep learning models for dental image analysis," *Scientific Reports*, vol. 11, no. 1, p. 6102, Mar. 2021, ISSN: 2045-2322. DOI: <http://dx.doi.org/10.1038/s41598-021-85454-5>. [Online]. Available: 10.1038/s41598-021-85454-5.
- [55] H. H. Lee et al., "Foundation models for biomedical image segmentation: A survey," *arXiv preprint arXiv:2401.07654*, 2024.
- [56] W. Lee, P. Jayathilake, Z. Tan, D. Le, H. Lee, and B. Khoo, "Muco-ciliary transport: Effect of mucus viscosity, cilia beat frequency and cilia density," *Computers & Fluids*, vol. 49, no. 1, pp. 214–221, 2011. DOI: <https://doi.org/10.1016/j.compfluid.2011.05.016>.
- [57] J. Leung, M. Rould, C. Konradt, C. Hunter, and G. Ward, "Disruption of tgphili alters specific parameters of toxoplasma gondii motility measured in a quantitative, three-dimensional live motility assay," *PloS one*, vol. 9, e85763, Jan. 2014. DOI: 10.1371/journal.pone.0085763.
- [58] F.-F. Li, R. Fergus, P. Perona, et al., "One-shot learning of object categories," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 594–611, 2006. DOI: <http://dx.doi.org/10.1109/TPAMI.2006.79>.
- [59] H. Li, H. Liu, D. Hu, J. Wang, and I. Oguz, "Promise: Prompt-driven 3d medical image segmentation using pretrained image foundation models," in *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, IEEE, 2024, pp. 1–5.
- [60] W. Liu et al., "Ssd: Single shot multibox detector," in *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, Springer, 2016, pp. 21–37.
- [61] X. Liu, Q. Xu, J. Ma, H. Jin, and Y. Zhang, "Mslrr: A unified multiscale low-rank representation for image segmentation," *IEEE transactions on image processing*, vol. 23, no. 5, pp. 2159–2167, 2014.
- [62] Y. Liu, S. J. Wagner, and T. Peng, "Multi-modality microscopy image style augmentation for nuclei segmentation," *Journal of Imaging*, vol. 8, no. 3, p. 71, 2022.
- [63] C. Lu, M. Marx, M. Zahid, C. W. Lo, C. Chennubhotla, and S. P. Quinn, "Stacked neural networks for end-to-end ciliary motion analysis," *arXiv preprint arXiv:1803.07534*, 2018. DOI: <https://doi.org/10.48550/arXiv.1803.07534>.

- [64] J. Ma and B. Wang, “Towards foundation models of biological image segmentation,” *Nature Methods*, vol. 20, no. 7, pp. 953–955, 2023.
- [65] J. Ma et al., “Segment anything in medical images and videos: Benchmark and deployment,” *arXiv preprint arXiv:2408.03322*, vol. 15, no. 1, p. 654, 2024.
- [66] A. Mahendran, J. Thewlis, and A. Vedaldi, “Cross pixel optical-flow similarity for self-supervised learning,” in *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part V 14*, Springer, 2019, pp. 99–116. DOI: <https://doi.org/10.48550/arXiv.1807.05636>.
- [67] L. Maier-Hein et al., “Why rankings of biomedical image analysis competitions should be interpreted with care,” *Nature communications*, vol. 9, no. 1, p. 5217, 2018.
- [68] M. Maška et al., “The cell tracking challenge: 10 years of objective benchmarking,” *Nature Methods*, vol. 20, no. 7, pp. 1010–1020, 2023.
- [69] M. A. Mazurowski, H. Dong, H. Gu, J. Yang, N. Konz, and Y. Zhang, “Segment anything model for medical image analysis: An experimental study,” *Medical Image Analysis*, vol. 89, p. 102 918, 2023.
- [70] E. G. Miller, N. E. Matsakis, and P. A. Viola, “Learning from one example through shared densities on transforms,” in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, IEEE, vol. 1, 2000, pp. 464–471. DOI: <https://doi.org/10.1109/CVPR.2000.855856>.
- [71] S. Na, Y. Guo, F. Jiang, H. Ma, and J. Huang, “Segment any cell: A sam-based auto-prompting fine-tuning framework for nuclei segmentation,” *arXiv preprint arXiv:2401.13220*, 2024.
- [72] H. Nam and B. Han, “Learning multi-domain convolutional neural networks for visual tracking,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4293–4302.
- [73] O. Oktay et al., “Attention u-net: Learning where to look for the pancreas,” *arXiv preprint arXiv:1804.03999*, 2018.
- [74] A. v. d. Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” *arXiv preprint arXiv:1807.03748*, 2018.
- [75] S. P. Quinn, M. J. Zahid, J. R. Durkin, R. J. Francis, C. W. Lo, and S. C. Chennubhotla, “Automated identification of abnormal respiratory ciliary motion in nasal biopsies,” *Science translational medicine*, vol. 7, no. 299, 299ra124–299ra124, 2015. DOI: <http://dx.doi.org/10.1126/scitranslmed.aaa1233>.
- [76] A. Radford et al., “Learning transferable visual models from natural language supervision,” in *International conference on machine learning*, PmLR, 2021, pp. 8748–8763.

- [77] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, “Transfusion: Understanding transfer learning for medical imaging,” in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32, Curran Associates, Inc., 2019. DOI: <https://doi.org/10.48550/arXiv.1902.07208>. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/eb1e78328c46506b46a4ac4a1e378b91-Paper.pdf.
- [78] M. Rakic, H. E. Wong, J. J. G. Ortiz, B. A. Cimini, J. V. Guttag, and A. V. Dalca, “Tyche: Stochastic in-context learning for medical image segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 11 159–11 173.
- [79] N. Ravi et al., “Sam 2: Segment anything in images and videos,” *arXiv preprint arXiv:2408.00714*, 2024.
- [80] F. Renard, S. Guedria, N. D. Palma, and N. Vuillerme, “Variability and reproducibility in deep learning for medical image segmentation,” *Scientific Reports*, vol. 10, no. 1, p. 13 724, 2020.
- [81] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings, part III 18*, Springer, 2015, pp. 234–241. DOI: 10.48550/arXiv.1505.04597.
- [82] C. Ryali et al., “Hiera: A hierarchical vision transformer without the bells-and-whistles,” in *International conference on machine learning*, PMLR, 2023, pp. 29 441–29 454.
- [83] G. Saadatnia and M. Golkar, “A review on human toxoplasmosis,” *Scandinavian journal of infectious diseases*, vol. 44, no. 11, pp. 805–814, 2012. DOI: 10.3109/00365548.2012.693197.
- [84] V. Sandfort, K. Yan, P. J. Pickhardt, and R. M. Summers, “Data augmentation using generative adversarial networks (cycleGAN) to improve generalizability in ct segmentation tasks,” *Scientific Reports*, vol. 9, no. 1, p. 16 884, Nov. 2019, ISSN: 2045-2322. DOI: <https://doi.org/10.1016/j.imu.2021.100779>.
- [85] T. H. Sanford et al., “Data augmentation and transfer learning to improve generalizability of an automated prostate segmentation model,” in *AJR Am J Roentgenol*, vol. 215, no. 6, pp. 1403–1410, Oct. 2020. DOI: <http://dx.doi.org/10.2214/AJR.19.22347>.
- [86] B. Settles, “Active learning literature survey,” 2009.
- [87] N. Sofroniew et al., “Napari: A multi-dimensional image viewer for python,” *Zenodo*, 2022.
- [88] N. Sofroniew et al., *napari: a multi-dimensional image viewer for Python*, version v0.4.16, If you use this software, please cite it using these metadata., May 2022. DOI: 10.5281/zenodo.6598542. [Online]. Available: <https://doi.org/10.5281/zenodo.6598542>.
- [89] C. Stringer, T. Wang, M. Michaelos, and M. Pachitariu, “Cellpose: A generalist algorithm for cellular segmentation,” *Nature methods*, vol. 18, no. 1, pp. 100–106, 2021. DOI: 10.1101/2020.02.02.931238.

- [90] M. Tan and Q. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” in *International conference on machine learning*, PMLR, 2019, pp. 6105–6114.
- [91] J.-Y. Tinevez et al., “Trackmate: An open and extensible platform for single-particle tracking,” *Methods*, vol. 115, pp. 80–90, 2017, Image Processing for Biologists, ISSN: 1046-2023. DOI: <https://doi.org/10.1016/j.ymeth.2016.09.016>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1046202316303346>.
- [92] S. A. Vaezi, G. Orlando, M. S. Fazli, G. E. Ward, S. N. Moreno, and S. Quinn, “A novel pipeline for cell instance segmentation, tracking and motility classification of toxoplasma gondii in 3d space,” in *SciPy*, 2022, pp. 60–63. DOI: <https://doi.org/10.25080/majora-212e5952-009>.
- [93] S. A. Vaezi and S. Quinn, “Training a supervised cilia segmentation model from self-supervision,” *Proceedings of the 23rd*, 2024.
- [94] D. A. Van Dyk and X.-L. Meng, “The art of data augmentation,” *Journal of Computational and Graphical Statistics*, vol. 10, no. 1, pp. 1–50, 2001. DOI: <http://dx.doi.org/10.1198/10618600152418584>.
- [95] J. E. Van Engelen and H. H. Hoos, “A survey on semi-supervised learning,” *Machine learning*, vol. 109, no. 2, pp. 373–440, 2020. DOI: <http://dx.doi.org/10.1007/s10994-019-05855-6>.
- [96] T. Vicar et al., “Cell segmentation methods for label-free contrast microscopy: Review and comprehensive comparison,” *BMC bioinformatics*, vol. 20, pp. 1–25, 2019.
- [97] S. Wang et al., “Annotation-efficient deep learning for automatic medical image segmentation,” *Nature communications*, vol. 12, no. 1, p. 5915, 2021.
- [98] M. Weigert, U. Schmidt, R. Haase, K. Sugawara, and G. Myers, “Star-convex polyhedra for 3d object detection and segmentation in microscopy,” in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE, Mar. 2020. DOI: 10.1109/wacv45572.2020.9093435. [Online]. Available: <https://doi.org/10.1109%2Fwacv45572.2020.9093435>.
- [99] C. Wen et al., “3DeeCellTracker, a deep learning-based pipeline for segmenting and tracking cells in 3D time lapse images,” *Elife*, vol. 10, Mar. 2021. DOI: 10.7554/eLife.59187. [Online]. Available: <https://doi.org/10.7554/eLife.59187>.
- [100] W. Weng and X. Zhu, “Inet: Convolutional networks for biomedical image segmentation,” *Ieee Access*, vol. 9, pp. 16 591–16 603, 2021.
- [101] A. Wolny et al., “Accurate and versatile 3d segmentation of plant tissues at cellular resolution,” *eLife*, vol. 9, C. S. Hardtke, D. C. Bergmann, D. C. Bergmann, and M. Graeff, Eds., e57613, Jul. 2020, ISSN: 2050-084X. DOI: 10.7554/eLife.57613. [Online]. Available: <https://doi.org/10.7554/eLife.57613>.
- [102] J. Wu et al., “Medical sam adapter: Adapting segment anything model for medical image segmentation,” *arXiv preprint arXiv:2304.12620*, 2023.

- [103] X. Xia and B. Kulis, “W-net: A deep model for fully unsupervised image segmentation,” *arXiv preprint arXiv:1711.08506*, 2017.
- [104] A. Yakimovich, A. Beaunon, Y. Huang, and E. Ozkirimli, “Labels in a haystack: Approaches beyond supervised learning in biomedical applications,” *Patterns*, vol. 2, no. 12, p. 100383, 2021, ISSN: 2666-3899. DOI: <https://doi.org/10.1016/j.patter.2021.100383>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2666389921002506>.
- [105] W. Yan et al., “Mri manufacturer shift and adaptation: Increasing the generalizability of deep learning segmentation for mr images acquired with different scanners,” *Radiology: Artificial Intelligence*, vol. 2, no. 4, e190195, 2020, PMID: 33937833. DOI: 10.1148/ryai.2020190195. eprint: 10.1148/ryai.2020190195. [Online]. Available: 10.1148/ryai.2020190195.
- [106] Z. Yan et al., “Biomedical sam 2: Segment anything in biomedical images and videos,” *arXiv preprint arXiv:2408.03286*, 2024.
- [107] M. Ye et al., “Hi-sam: Marrying segment anything model for hierarchical text segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
- [108] X. Yi, E. Walia, and P. Babyn, “Generative adversarial network in medical imaging: A review,” *Medical image analysis*, vol. 58, p. 101552, 2019. DOI: <https://doi.org/10.48550/arXiv.1809.07294>.
- [109] M. Zain, E. Miller, S. Quinn, and C. Lo, “Low level feature extraction for cilia segmentation,” in *Proceedings of the Python in Science Conference*, 2022. DOI: <https://doi.org/10.25080/majora-212e5952-026>.
- [110] M. Zain et al., “Towards an unsupervised spatiotemporal representation of cilia video using a modular generative pipeline,” in *Proceedings of the Python in Science Conference*, 2020. DOI: <http://dx.doi.org/10.25080/Majora-342d178e-017>.
- [111] S. Zhang et al., “Biomedclip: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs,” *arXiv preprint arXiv:2303.00915*, 2023.
- [112] W. Zhang, H. Cheng, and J. Gan, “Munet: A multi-scale u-net framework for medical image segmentation,” in *2020 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2020, pp. 1–7.
- [113] Y. Zhang et al., “Evf-sam: Early vision-language fusion for text-prompted segment anything model,” *arXiv preprint arXiv:2406.20076*, 2024.
- [114] Z. Zhao et al., “One model to rule them all: Towards universal segmentation for medical images with text prompts,” *arXiv preprint arXiv:2312.17183*, 2023.
- [115] T. Zhou, S. Ruan, and S. Canu, “A review: Deep learning for medical image segmentation using multi-modality fusion,” *Array*, vol. 3, p. 100004, 2019.