

A Machine Learning Model for Cellular Microscopy Segmentation

Undergraduate Thesis by Emilio Aponte

April 21, 2025

Dr. Craig Forest, Advisor
Woodruff School of Mechanical Engineering
Georgia Institute of Technology

Dr. Danfei Xu, Second Reader
School of Interactive Computing
Georgia Institute of Technology

Contents

1	Introduction	3
1.1	Cell Segmentation	3
1.2	Transfer Learning and the Segment Anything Model (SAM)	3
1.3	Research Goals	4
1.4	Challenges	4
2	Methods and Materials	4
2.1	Morphology descriptors	4
2.2	Collect cell features	6
2.3	Clustering	6
3	Results	6
3.1	Data Cleaning	6
3.2	k -Means Results	8
3.3	DBSCAN Results	8
4	Discussion	9
4.1	SAMCell Continued Development	9
4.2	Cell Clustering	9

1 Introduction

1.1 Cell Segmentation

Image analysis has revolutionized medical research and clinical practice by providing sophisticated tools for the interpretation and quantification of medical images. Techniques such as MRI, CT scans, and microscopy generate vast amounts of visual data, which require accurate and efficient analysis to support diagnostics, treatment planning, and biomedical research. The application of deep learning and computer vision has significantly enhanced the ability to automatically process and analyze these images, leading to improved detection of diseases, tracking of disease progression, and evaluation of treatment responses. The deep learning architecture known as U-Net has previously been used for segmentation in medical microscopy, and its more advanced cousins V-Net and Progressive Dense V-Net for MRI and CT data, respectively [10]. Advanced image segmentation models, in particular, play a crucial role in delineating anatomical structures, identifying pathological regions, and enabling precise quantitative analysis in various medical imaging tasks.

Cell segmentation, the process of identifying and delineating individual cells within microscopy images, is a fundamental task in biomedical research. Accurate cell segmentation is essential for a range of applications, including cell counting, measuring morphological features, tracking cell movements, and analyzing cell-cell interactions. These tasks are critical for understanding cellular behaviors, disease mechanisms, and the effects of therapeutic interventions at the cellular level. For instance, in cancer research, precise segmentation of tumor cells can provide insights into tumor growth patterns and metastatic potential. Due to a lack of options that are both accurate and convenient, many biomedical researchers still opt to perform these tasks manually, which places a significant strain on the speed at which research analysis can take place.

Techniques for image segmentation have evolved over the last decades. The first approaches used pixel intensity thresholds and statistical methods to extract separate image regions [17]. This only works on processed high-contrast images, and is not yet suitable for application in cell microscopy. Since then, new approaches have been developed, and with the advent of advanced deep neural networks in recent years, cell segmentation has become more promising. Encoder-Decoder models compress and decode image data in such a way that semantic features are revealed to allow image segmentation to work [10]. In the realm of deep learning, several types of convolutional models now exist. Convolutional networks are powerful in that applying convolutions in sequence has the effect of extracting human-recognizable image features. The convolution layers are fine-tuned to recognize certain visual features with segmented training images, and upon inference they show promising image segmentation results [10]. In order to work for the highly specialized task of cell segmentation, a few additional challenges remain.

1.2 Transfer Learning and the Segment Anything Model (SAM)

The Segment Anything Model (SAM), developed by Meta, represents a significant advancement in the field of image segmentation. SAM is designed to be a versatile and powerful model capable of identifying and segmenting objects within an image, irrespective of the context. This adaptability makes SAM a valuable tool for various image analysis tasks, including those in the medical field. However, the direct application of SAM to medical images, such as those from cellular microscopy, has revealed limitations in its accuracy and performance, particularly in tasks requiring fine-grained segmentation of densely packed cells.

To address these challenges, several adaptations and specialized models have been developed. For example, SAMed and SAM-IE are tailored versions of SAM that aim to improve its performance in medical image segmentation. These models leverage additional training on medical datasets and incorporate domain-specific optimizations. Other approaches, such as the star-convex polygons method and its extensions like StarDist and StarDist-3D, offer alternative techniques for cell segmentation, focusing on accurate shape representation and segmentation of crowded cells. The development and refinement of these models are crucial for advancing the state of medical image analysis, enabling more precise and efficient processing of complex medical imagery.

One such specialized models, developed at the Georgia Tech Precision Bio System Lab, is SAMCell. A SAM model trained on a curated dataset of annotated cell microscopy images and a wide range of cell types. The model has shown promising improvements in the original SAM's

ability to distinguish individual cells out of clumps, in a level comparable to an experienced biologist. Importantly, SAMCell is able to accurately segment images of cells out of a variety of cell types and shapes thanks to an approach similar to StarDist. SAMCell uses its trained model to identify individual cell centers and then performs a pixel search to predict which areas of the image match to each corresponding cell.

1.3 Research Goals

Despite the number of new options biologists have to automate the task of cell segmentation, virtually none have caught on and the majority of biomedical research labs rely on human annotators to collect and analyze microscopy data. The goal of this paper is to facilitate the process of cell segmentation for researchers by providing a simple solution. We aim to build upon the advancements of SAMCell and increase its usability for common use in biomedical research laboratories through the improvement of the model and development of an accessible lightweight User Interface for users to make use of SAMCell in a manner that minimally disrupts lab procedure and resources.

A modern challenge of cell culture analysis is a lack of standard metrics. Scientists with experience in cell research form intuitions regarding cell culture analysis, but their ability to confirm results are limited. If this project is successful in attracting users in biomedical research, the primary goal of this project is to leverage SAMCell, and this new ability to extract concrete cell masks, to standardize the methods by which certain cell culture metrics are obtained.

1.4 Challenges

With the primary goal of developing a remote interface for SAMCell, we foresee a few technical challenges. The first is in the scaling of the application. The hardware requirements for the model are such that with a sufficient amount of concurrent users, the runtime of SAMCell can become impractical. We will need to reach a scaling plan that allows the endpoint to meet demand. This may require a re-evaluating of the pricing model, as hosting SAMCell long-term will incur costs.

Cell microscopy images are high resolution images that can become difficult to transmit online. An investigation can be done into SAMCell’s performance on low resolution images. It may be possible to reduce image quality for transmission without loss of output accuracy.

2 Methods and Materials

Without trusted methods for cell segmentation at a large scale, cell culturing research is largely subjective and dependent on the individual researcher’s discretion. Research assistants have to make educated predictions when it comes to analyzing an image and counting the cells. This analysis process is also limited to determining the general location of cells in an image. The ability to determine the shape of cells within a cell culture has never been available, so there has been no need to standardize the morphology characterization of cell cultures.

Our SAMCell model shows promise in introducing a convenient method for shape extraction across large microscopy datasets. If successful, biologists will require such a standard. We propose a set of morphology descriptors that show high correlation with cell characterization as a start to establish a cell morphology analysis standard.

We conducted a unsupervised clustering on a sample set of cells to determine the effectiveness of our proposed morphology features. These features were inspired from various common morphology metrics, such as river rock morphology and metal grain classification in metallurgy. We generate 443 Human Embryonic Kidney (HEK) cell masks using SAMCell and run our morphology algorithms to generate features for each individual data point.

2.1 Morphology descriptors

Through a process of literature review and performance optimization we develop algorithms to determine the following features given an individual segmented cell mask. Some of these are inter-dependent. There may be several adequate definitions for some of the mathematical properties of the given shape, so we consider as many as possible and determine the best one.

1. Cell center. Considering the cell mask as a uniform shape, there are several centers to consider. Determining the cell center allows us to calculate other cell features such as the minimum/maximum radius, or establishing a radial profile.
 - (a) Chebyshev center. The point which is closest to its furthest edge point, or alternatively the point which is furthest from its closest edge point.

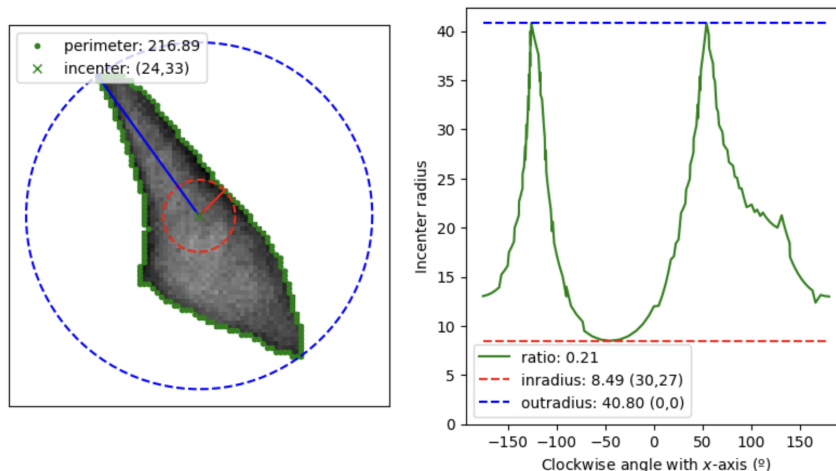


Figure 1: An example of the Chebyshev center (left). For each pixel coordinate in the cell mask we determine the (Euclidean) distance to its furthest border pixel and pick the point for which this distance is minimal. Determining the center allows further analysis to be done on the cell, such as establishing a radial profile (right).

- (b) Inscribed circle center. The center of the circle with maximal area contained entirely within the shape. Alternatively the circumscribed circle center, the center of the circle with minimal area which entirely contains the shape. Finding the true inscribed or circumscribed circle is an optimization problem that requires linear programming approaches with unreasonable computational complexity for our applications.
 - (c) Centroid. The mean x and y coordinates of all points in the shape. The centroid is a poor choice for applications is cell morphology due to the prevalence of edge cases such as sickle-shaped cells.
2. Cell area and perimeter. We want to find the area and border perimeter (relative to the image - pixels) of the shape. We order the border pixels by the angle they make with the chosen center and sum adjacent pixel distances.
3. Radial Profile. Having defined a cell center, we can generally trace out a continuous curve describing the cell's border. Analysis and comparison of these curves can be used to characterize shapes.
4. Inscribed and circumscribed circles. An optimization problem of maximizing or minimizing circle area with some constraints. As with the radial profile, it may be used to characterize a shape's border. Since the circles need not be aligned, the discrepancy between circle centers may also be used as a shape descriptor.

- Length of major and minor axes. The major axis is the straight line between the two furthest points on the cell's perimeter. The minor axis is perpendicular to the major axis such that they intersect at the cell's center (depends on which center we chose to use). Characterizing the oblong qualities of the cell can give information about its type, health, or life stage.

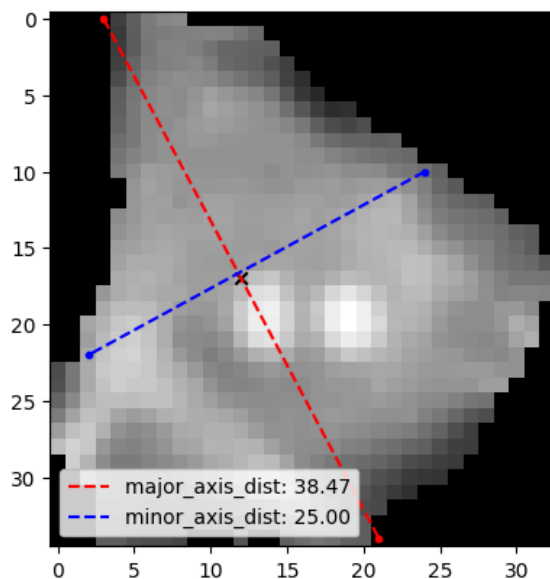


Figure 2: An example cell and its corresponding major axis (red) and minor axis (blue).

2.2 Collect cell features

Once the cell morphology features were determined we obtained a bank of 443 cell masks by segmenting an image of HEK 293 cells from the Georgia Tech Precision Biosystems Laboratory using SAMCell. We then computed the chosen morphology features for each cell to construct a dataset of individual cells and a low-level description of their shape.

2.3 Clustering

After cleaning outlying cells, most of which are likely a result of SAMCell errors of the kind previously identified (section 1.3), we attempt to use this data to learn to potentially classify HEK cells throughout their distinct developmental phases from their cell masks. We implement two unsupervised clustering methods in an attempt to identify patterns in the way cells share morphological qualities with other cells.

- k -Means Clustering [11]
- Density Based Spatial Clustering of Applications with Noise (DBSCAN) [1]

3 Results

3.1 Data Cleaning

Having the morphology dataset we can employ unsupervised learning to classify cells in a parameter space of their morphology metrics. We implement various clustering methods from the data points, each cell having 2-dimensional data, that being the cell's perimeter and the length of its major axis. Clustering methods such as those we showcase, k -Means and DBSCAN, can be applied to multidimensional data points. As we intend to develop further morphology metrics, this approach has the potential to be even more useful.

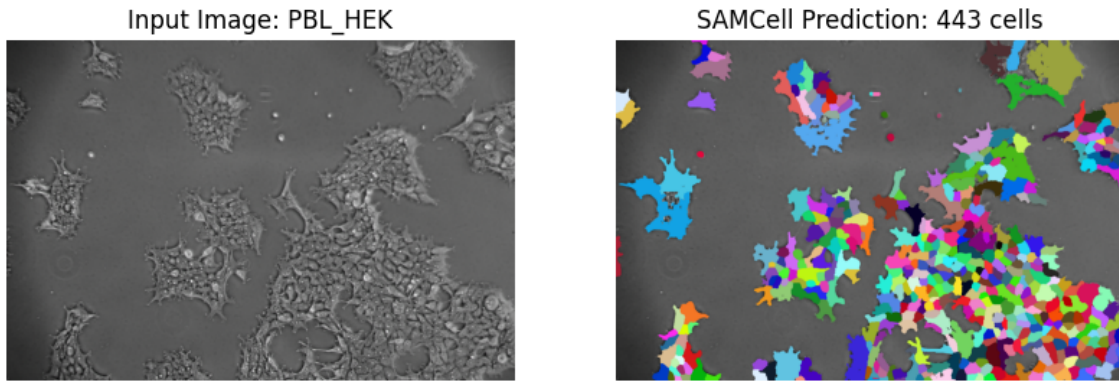
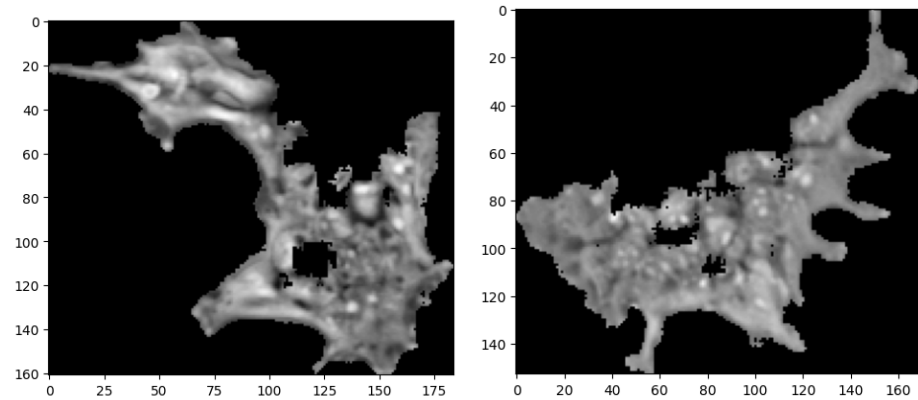


Figure 3: (Left) A raw microscopy image from PBL’s HEK dataset. It contains a mix of singular and clumped cells to test the limitations of SAMCell. (Right) The image’s SAMCell predicted output, with cell masks assigned a unique random color.

As seen in Figure 3, SAMCell identified 443 distinct HEK cells. We ran our selected morphology metrics to find the (1) Chebyshev center coordinates, (2) cell perimeter (in pixel units), and (3) the length of major axis (in pixel units). This exercise reveals some minor flaws in the SAMCell model. 25 of the 443 reported cells had been segmented with empty masks. After removing these data points, we find two notable outliers due to their unusually high perimeter. After investigation, the following cell masks were found.



(a) Outlier cell mask, cell 286 with perimeter 22178.61 (pixel units). (b) Cell 438 with perimeter 20290.12 (pixel units).

Figure 4: Two cells with abnormally large perimeters. Investigation clearly reveals these to be incorrectly segmented cell clumps. This is one of SAMCell’s most well-documented limitations.

3.2 k -Means Results

With outliers removed, we ran three k -Means experiments, each with 1000 maximum iterations and relative loss threshold of 10^{-5} , sensible stopping criteria for convergence.

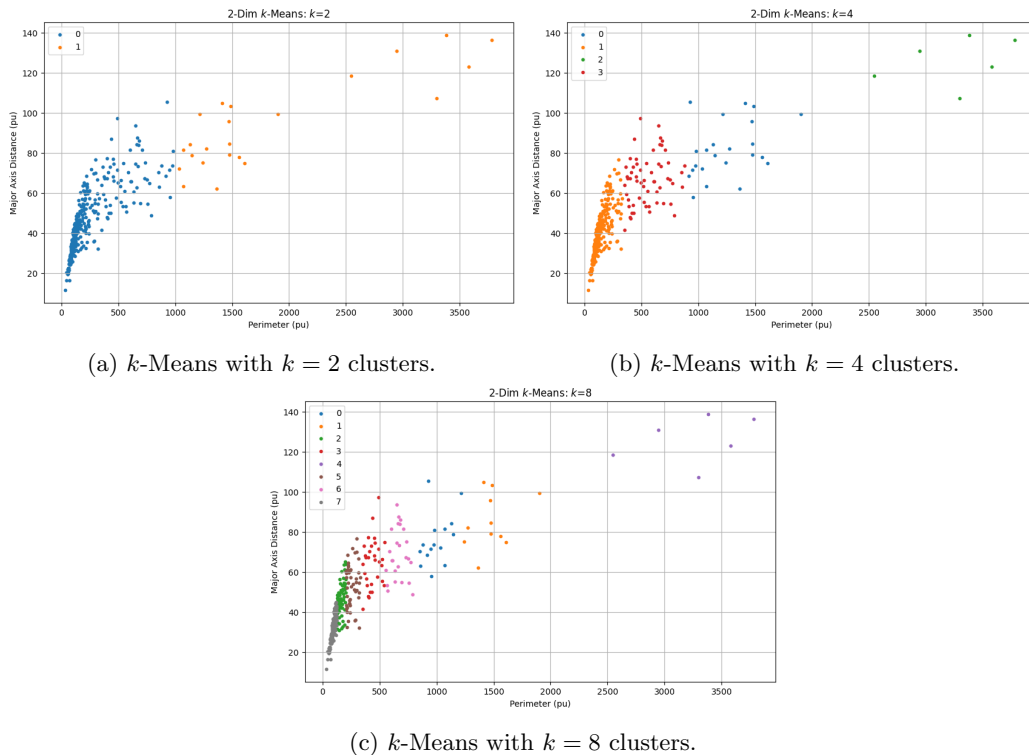


Figure 5: 2-Dimensional k -Means clustering of our segmented HEK cells.

3.3 DBSCAN Results

We also ran the DBSCAN procedure on the same 2-dimensional data, with search radius $\epsilon = 50$ and the standard minimum points $D + 1 = 3$, where $D = 2$ is the data dimensionality.

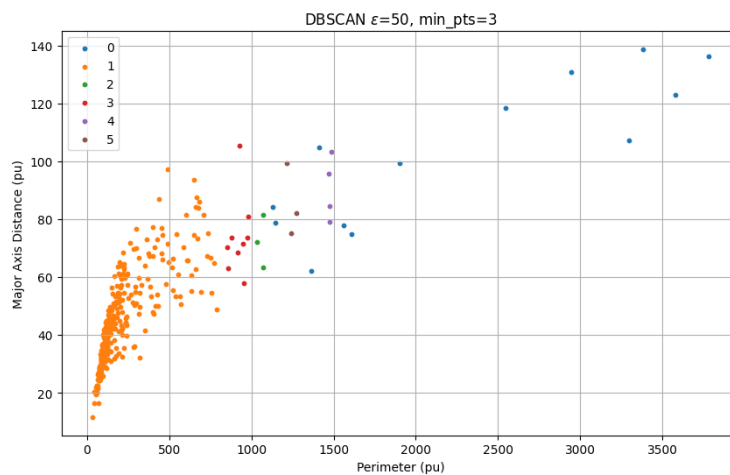


Figure 6: 2-Dimensional DBSCAN reveals 6 natural clusters.

4 Discussion

4.1 SAMCell Continued Development

Having developed an integrated platform for biologists to interact with our SAMCell model will have several implications for the field as well as the future of the SAMCell project. When new users use the program, we will gain two valuable pieces of data. The first being a brand new set of images that SAMCell has never seen during training, and the second is an expert-validated annotation of the image. We hope to be able to use this data to increase the training sample that SAMCell can use for future training.

We anticipate significant improvements in the quality and scope of biomedical research that will be achieved by SAMCell users, particularly through its simple and accessible new interface. Anecdotal accounts show the features implemented are popular among current users, and we anticipate a growth in the widespread use of the model. The ability to instantaneously count the cells in an image and then re-annotate the output to correct minor mistakes will significantly reduce the overhead in biology research methods. Additionally, the ability to obtain a cell's shape is data that is hardly ever collected and SAMCell can determine in a matter of minutes. Paired with our morphology metrics, we anticipate significant new insights related to cell morphology in a wide range of medical fields.

4.2 Cell Clustering

We provided an example analysis of a HEK 293 cell culture informed by our cell morphology features. With similar approaches, leveraging the advent of our accurate segmentation method, similar investigations can be conducted by experts to learn more about a variety of cell types. This has the potential to advance a number of important investigations in modern biology.

This initial investigation into the uses of our morphology features also revealed limitations of the current iteration of SAMCell. We found outliers by calculating the perimeter and major axis of each segmented cell. Computationally lightweight features such as this could be implemented in the SAMCell interface to provide real-time feedback for researchers. The interface can use these features to automatically identify potentially erroneous segmentations.

We intend to continue development of computable morphology metrics to be added potentially as features in the SAMCell interface. The hope is to develop informed metrics that can be adopted and standardized by the cell research community. Immediate goals include finding a robust cell center. In the case of abnormally shaped cells, the Chebyshev center is not ideal as there may be multiple valid centers in some cases, and none at all in others. The inscribed circle approach show more promise, but would require fast optimization in order to be applied to large scale segmented data sets. Not discussed in this paper are features that take into account additional information to the cell mask. We may consider relative intensity and texture features that use the color information in microscopy images to characterize a cells health and shape not in the plane in which the image is taken.

References

- [1] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In Evangelos Simoudis, Jiawei Han, and Usama M. Fayyad, editors, *KDD*, pages 226–231. AAAI Press, 1996.
- [2] Marc Fischer, Alexander Bartler, and Bin Yang. Prompt tuning for parameter-efficient medical image segmentation. *Medical Image Analysis*, 91:103024, 2024.
- [3] Sunwoo Han, Khamson Phasouk, Jia Zhu, and Youyi Fong. Optimizing deep learning-based segmentation of densely packed cells using cell surface markers. *BMC Medical Informatics and Decision Making*, 2024.
- [4] Sheng He, Rina Bao, Jingpeng Li, Jeffrey Stout, Atle Bjornerud, P. Ellen Grant, and Yangming Ou. Computer-vision benchmark segment-anything model (sam) in medical images: Accuracy in 12 datasets, 2023.
- [5] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023.
- [6] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.
- [7] Yan Liu, Maojun Zhang, Zhiwei Zhong, and Xiangrong Zeng. A novel adaptive cubic quasi-newton optimizer for deep learning based medical image analysis tasks, validated on detection of covid-19 and segmentation for covid-19 lung infection, liver tumor, and optic disc/cup. *Medical Physics*, 50(3):1528–1538, 2023.
- [8] Francesco Marzola, Nens van Alfen, Jonne Doorduyn, and Kristen M. Meiburger. Deep learning segmentation of transverse musculoskeletal ultrasound images for neuromuscular disease assessment. *Computers in Biology and Medicine*, 135:104623, 2021.
- [9] Maciej A. Mazurowski, Haoyu Dong, Hanxue Gu, Jichen Yang, Nicholas Konz, and Yixin Zhang. Segment anything model for medical image analysis: An experimental study. *Medical Image Analysis*, 89:102918, 2023.
- [10] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3523–3542, 2022.
- [11] Dan Pelleg and Andrew Moore. Accelerating exact k-means algorithms with geometric reasoning. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '99, page 277–281, New York, NY, USA, 1999. Association for Computing Machinery.
- [12] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.
- [13] Uwe Schmidt, Martin Weigert, Coleman Broaddus, and Gene Myers. Cell detection with star-convex polygons. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2018 - 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II*, pages 265–273, 2018.
- [14] Changyan Wang, Haobo Chen, Xin Zhou, Meng Wang, and Qi Zhang. Sam-ie: Sam-based image enhancement for facilitating medical image diagnosis with segmentation foundation model. *Expert Systems with Applications*, 249:123795, 2024.
- [15] Martin Weigert and Uwe Schmidt. Nuclei instance segmentation and classification in histopathology images with stardist. In *The IEEE International Symposium on Biomedical Imaging Challenges (ISBIC)*, 2022.

- [16] Martin Weigert, Uwe Schmidt, Robert Haase, Ko Sugawara, and Gene Myers. Star-convex polyhedra for 3d object detection and segmentation in microscopy. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2020.
- [17] S.D. Yanowitz and A.M. Bruckstein. A new method for image segmentation. *Computer Vision, Graphics, and Image Processing*, 46(1):82–95, 1989.
- [18] Kaidong Zhang and Dong Liu. Customized segment anything model for medical image segmentation, 2023.
- [19] Yichi Zhang, Zhenrong Shen, and Rushi Jiao. Segment anything model for medical image segmentation: Current applications and future directions. *Computers in Biology and Medicine*, 171:108238, 2024.