Zurich University of Applied Sciences

Department School of Engineering

Centre for Artificial Intelligence

BACHELOR THESIS

Deep-Learning-Based Cell Segmentation for the Detection of Thyroid Cancer in Single Cells

---

Authors:
Tenzin Samdrup LANGDUN
Martin OSWALD

Supervisors:
Prof. Dr. Thilo STADELMANN
Prof. Dr. Anna N. YAROSLAVSKY

Submitted on
June 9, 2023

Study program:
Computer Science

Declaration of Authorship

We, Tenzin Samdrup LANGDUN, Martin OSWALD, declare that this thesis titled, "Deep-Learning-Based Cell Segmentation for the Detection of Thyroid Cancer in Single Cells" and the work presented in it are our own. We confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where we have consulted the published work of others, this is always clearly attributed.
- Where we have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely our own work.
- We have acknowledged all main sources of help.
- Where the thesis is based on work done by ourselves jointly with others, we have made clear exactly what was done by others and what we have contributed ourselves.

Signed: 

Date: June 9th, 2023

ZURICH UNIVERSITY OF APPLIED SCIENCES

Abstract

Centre for Artificial Intelligence
School of Engineering

Bachelor of Science

Deep-Learning-Based Cell Segmentation for the Detection of Thyroid Cancer in Single Cells

by Tenzin Samdrup LANGDUN, Martin OSWALD

Automated segmentation of biological images plays a crucial role in accelerating the annotation process and can greatly aid in the diagnosis of various medical conditions. In the context of thyroid cancer detection, the time-consuming manual segmentation of fluorescence polarization (Fpol) images using methylene blue-stained cells poses a challenge, hindering widespread implementation. To address this limitation and provide a cost-effective alternative, this paper focuses on the development of a U-Net-based deep learning model for automated cell segmentation. The approach addresses the limitations of manual segmentation and the inherent ambiguity in human-annotated images. The incorporation of non-ambiguous labeled images significantly enhances the model's performance. The research also delves into the concept of diminishing returns when adding these images and explores the potential to reduce manual labor through the application of Semi-Supervised Active Learning (SSAL), where the model is bootstrapped using pseudo-labeled images. Initially, training the model with additional pseudo-labeled images led to a decline in segmentation performance. However, manual correction of these pseudo-labeled images resulted in a slight improvement in performance and a substantial reduction in annotation time, with an average saving of 65%. By deploying this automated approach, the time required for manual segmentation and calculation of Fpol values could be reduced by over 99.7%. This reduction highlights the transformative potential of this SSAL framework in revolutionizing thyroid cancer diagnostics.

Keywords: cell segmentation, partially labeled data, thyroid cancer diagnosis, semi-supervised active learning, methylene blue polarization

Acknowledgements

We would like to thank Prof. Thilo Stadelmann, Dr. Anna Yaroslavsky, Dr. Peter Jermain and Dr. Ahmed Abdulkadir for their valuable guidance and support throughout the project. Would also like to thank Santana Wright for assisting us with the evaluation of our model. We are very grateful to be given the opportunity to work on such an important and cutting-edge project.

Contents

List of Figures

List of Tables

List of Abbreviations

CEAL	**C**ost-**E**ffective **A**ctive **L**earning
CNN	**C**onvolutional **N**eural **N**etwork
CV	**C**omputer **V**ision
DL	**D**eep **L**earning
Fpol	**F**luorescent **Pol**arization
FNAC	**F**ine **N**eedle **A**spiration **C**ytology
MB	**M**ethylene **B**lue
MBFP	**M**ethylene **B**lue **F**luorescent **P**olarization
ML	**M**achine **L**earning
NN	**N**eural **N**etwork
SSAL	**S**emi-**S**upervised **A**ctive **L**earning
SSL	**S**emi-**S**upervised **L**earning

1 Introduction

1.1 Motivation

Despite the high prevalence of thyroid cancer, the currently available diagnostic procedures are plagued by inaccuracies, high costs, and potential risks to patients [1]. The American Cancer Society estimates that over 43,000 cases of thyroid cancer are diagnosed in the US every year, making it the predominant cancer type affecting the endocrine system [2]. More than half of the adult population will experience a thyroid lump, with approximately 5% of those cases proving to be malignant [1].

Fine needle aspiration cytology (FNAC) is the standard of care for detecting and testing thyroid nodules [1], [3]. However, the current analysis of FNACs yields indeterminate results in nearly 30% of cases, and the overall diagnostic accuracy ranges from 60.2% to 68.8% [3], [4].

To verify the accuracy of FNAC, additional diagnostic procedures such as histopathological examination are often performed [5]. However, these procedures may be invasive, costly, and time-consuming, adding further burden and potential risks to patients [5].

Options available to reduce the diagnostic uncertainty are either more invasive, costly, or both [6]. One of these options, surgical biopsy, also known as surgical resection, involves removing part or all of the thyroid gland to examine the tissue for signs of cancer. This method may lead to postoperative hypocalcemia [7], which is characterized by low levels of calcium in the blood following the surgery. In rare cases, it may also lead to other disorders, such as vocal cord paralysis [7]. In cases where the FNAC biopsy results are inconclusive, molecular testing can be performed on the biopsy sample [6]. Molecular genetic testing is expensive, requires a dedicated sample without a guarantee of a definitive diagnosis, and has long turnaround times [6]. Therefore, the development of a more accurate and cost-effective method for thyroid cancer diagnosis could benefit patients by reducing the need for costly and invasive procedures.

1.2 Problem Statement & Work Outline

The research laboratory led by Professor Yaroslavsky at UMass Lowell has developed a cutting-edge technology for the detection of cancer at the cellular level [3]. This innovative approach utilizes fluorescence polarization (Fpol) signals obtained from cells that have been stained with methylene blue (MB). The method holds the potential to decrease unnecessary diagnostic surgeries and expensive molecular analyses, thereby saving significant amounts of time and resources for the healthcare system and substantially reducing morbidity associated with unwarranted surgery [3].

However, a major impediment to the efficiency of this method is the need for manual cell segmentation in the images for subsequent analysis. This process is labor-intensive, time-consuming, and dependent on the user's skills [3]. If a robust automated cell segmentation

algorithm could be developed, it would enable rapid assessments of FNAC samples to identify and classify cancer cells at a lower cost.

The objective of this thesis is to develop such an automatic cell segmentation method, leveraging deep learning for semantic segmentation in computer vision, such as the U-Net architecture. The main challenge addressed in our thesis is the ambiguity in the manually segmented images provided by the human annotators: These images may contain cells that were not segmented by the annotators but can still be classified as "eligible cells" (discussed in Chapter 3.1) - their labeling is hence incomplete. Insufficient or partial labeling of training images poses a significant obstacle for any machine learning system aiming to learn solely from human annotations, as it becomes challenging to determine what should be segmented and what should be left unsegmented. It is important to note that the goal is not to segment every single cell, but those that match specific criteria established by the experts (see Appendix A.1).

We therefore pursue two goals. First, we want to test the hypothesis whether the inclusion of non-ambiguously, fully labeled images (see Section 3.1) in the training set will enhance the model's performance (first objective). If it proves correct, we will then investigate the relationship between the added number of non-ambiguously labeled images and the improvement of the model's results (second objective). Finally, we are also seeking to find a balance between manual labor and automated labeling, i.e. a "sweet spot" that optimizes cost and reward. Manual labeling of images is a time-consuming process, and finding a way to reduce this labor while improving the accuracy of the model is a key objective of our work. Therefore, we are interested in exploring if the model can be bootstrapped to pseudo-label images, thereby reducing the need for human manual labor (third objective). By achieving these goals, we excpect to substantially improve the efficiency and accuracy of thyroid cancer diagnosis, potentially leading to a decrease in the disease's mortality rate.

This paper is structured as follows: Chapter 2 presents previous research and the technologies that are relevant for the experiments. Chapter 3 introduces the data and methods that were the basis for the experiments and results presented in Chapter 4. The final conclusion summarizes the value of the present study in the context of current research trends and gives an outlook towards the future directions the field could take.

2 Foundations & Related Work

This section provides a comprehensive overview of the foundational knowledge and related works that form the basis of our research. We begin by examining the technique of methylene blue fluorescence polarization and its relevance to our objectives (Section 2.1). We then explore the intersection of computer vision and deep learning (Section 2.2), followed by an exploration of image segmentation techniques (Section 2.3). The U-Net architecture, a framework for biomedical image segmentation, is introduced in Section 2.4. Additionally, we discuss strategies for leveraging small datasets in deep learning (Section 2.5) and approaches for handling partially labeled data (Section 2.6). These concepts provide the necessary background for understanding our proposed methodology and the experimental results presented in the subsequent chapters.

2.1 Methylene Blue Fluorescence Polarization

Dr. Anna Yaroslavsky, Dr. Peter Jermain, and their team have addressed the challenge of accurately diagnosing thyroid cancer using a needle biopsy technique [3]. They developed a method that employs a special kind of imaging, called confocal fluorescence polarization imaging, using a dye called methylene blue to detect cancerous cells in human samples. Their research suggests that this imaging method can provide a reliable way to identify thyroid cancer, potentially transforming the way cancer diagnosis is done at the cellular level from a subjective visual analysis to a more objective and precise measurement [3].

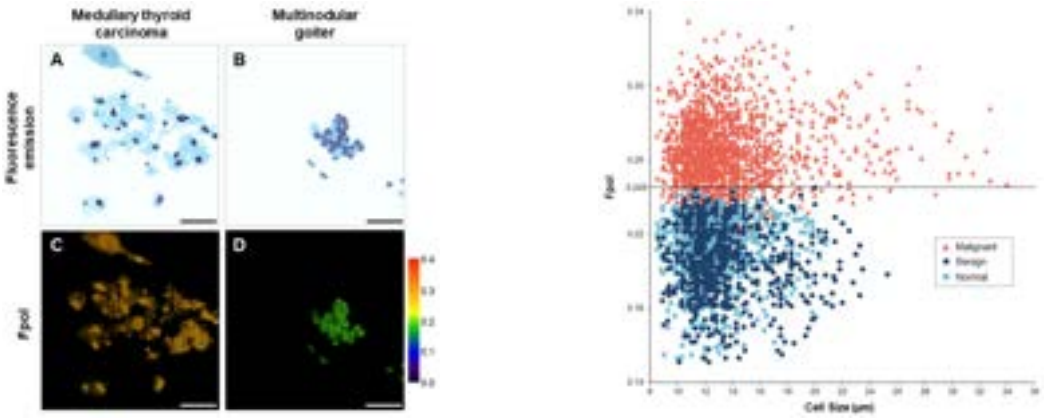
The technique utilizes exogenous fluorescence polarization (Fpol) of methylene blue (MB) as a quantitative marker for thyroid cancer. Fpol, short for fluorescence polarization, measures the degree of alignment of fluorophores (in this case, MB) in the sample. Methylene blue is a fluorescent dye that selectively binds to cancerous cells and exhibits higher fluorescence polarization when bound to cancer cells than to normal or benign cells. The reason for the enhanced fluorescence polarization of MB is its greater concentration within the mitochondria of cancer cells, along with a reduced fluorescence lifetime compared to normal cells [8].

The results of the study revealed that MBFP can effectively differentiate between cancerous and non-cancerous thyroid cells. Notably, MBFP exhibited significantly higher values in papillary carcinoma and follicular carcinoma samples compared to normal or benign cells, as illustrated in Figure 2.1a [3].

MBFP imaging, as depicted in Figure 2.1b, effectively distinguished between benign and malignant thyroid nodules. The results demonstrated significantly higher Fpol values in medullary thyroid carcinoma compared to multinodular goiter. Notably, benign and normal cells exhibited Fpol values lower than 0.245 [3]. These findings highlight the potential of MBFP as a promising alternative to current diagnostic methods in thyroid nodule classification. Compared to other imaging techniques, MBFP imaging is cost-effective, fast, and easy to interpret. It offers high contrast images of cancerous cells and does not require complex data processing or evaluation of cell morphology [3]. The findings suggest that MBFP can serve as a quantitative

biomarker for thyroid cancer, complementing traditional visual examination. Digital staining algorithms can mimic cytological stains, aiding pathologists in interpreting the images.

Each sample required approximately 10 minutes for image acquisition. However, the most time-consuming step in data processing was manual cell segmentation, which took about 10-15 minutes per sample for a trained expert [3]. A decrease in segmentation time may enable rapid on-site cancer detection.



(a) (b)

FIGURE 2.1: (a) Fpol values of medullary thyroid carcinoma and multinodular goiter (b) Benign and normal cells show values less than 0.245 [3].

2.2 Deep Learning

In the context of cell segmentation, both computer vision (CV) and deep learning play crucial roles. Traditional CV techniques, such as thresholding and morphological filtering, have been employed to segment cells in microscopy images. However, these techniques often require manual tuning of parameters and are sensitive to variations in image quality, cell morphology, and staining [9]. Deep learning techniques have the potential to overcome these limitations by automatically learning features and representations from the data, leading to more accurate and robust cell segmentation algorithms [10].

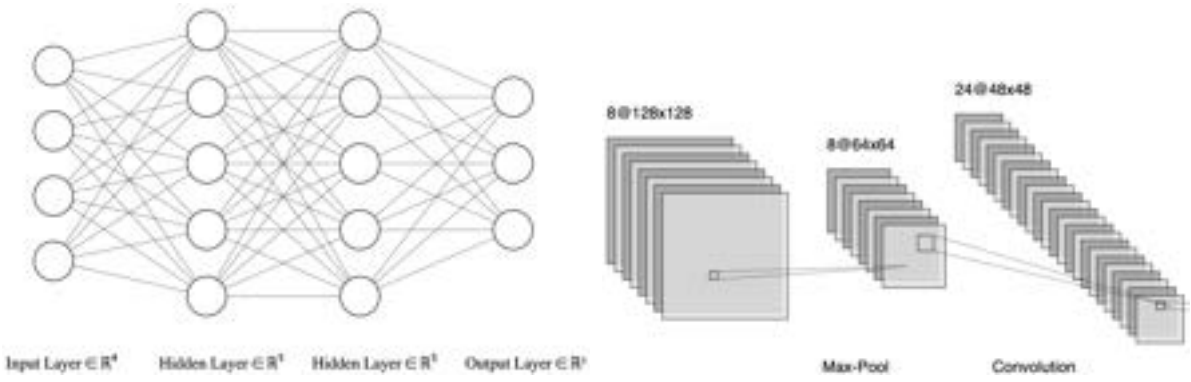
DL is a subset of machine learning (ML) that focuses on learning data representations at multiple abstraction levels. DL models, unlike traditional ML, automatically learn complex and abstract features from raw data by leveraging interconnected layers [11]. This eliminates the need for manual feature engineering, which is typically required in traditional ML approaches.

Neural networks (NNs) form the basis of deep learning, was inspired by biological neurons with afferent and efferent connections that are modelled as interconnected layers of nodes or "neurons" that map inputs to outputs using an internal state [11]. A feed forward network, as depicted in Figure 2.2a, demonstrates this concept, with neurons receiving inputs, performing computations, and passing results to subsequent layers [12].

Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) are a specialized kind of neural network designed for processing data with a grid-like topology, such as images [11]. CNNs are particularly effective for tasks related to image recognition and classification, utilizing various types of layers such

as convolutional, pooling, and fully connected layers, among others. [11] The architecture of a CNN is designed to take advantage of the grid-like structure of an input image by utilizing kernels to extract data (pixels) that are next to each other. Figure 2.2b illustrates the architecture of a CNN where the kernel extracts local information [11], [13].



(a) (b)

FIGURE 2.2: (a) Feed Forward Neural Network (FFNN) (b) Convolutional Neural Network (CNN) [14]

Training Deep Learning Models

Training deep learning models involves adjusting the weights of the network to minimize the difference between the predicted and actual outputs. During training, the input is passed forward through the network to generate a prediction. The prediction is then compared to the actual output, and the error is propagated back through the network to adjust the weights. This process is repeated until the network's predictions are satisfactory [13].

One of the challenges in training deep learning models is maintaining a balance between underfitting and overfitting. Underfitting (Figure 2.3a) occurs when the model is too simple to capture the complexity of the data, resulting in poor performance on the test data. Overfitting (Figure 2.3c), on the other hand, occurs when the model learns the training data too well, including its noise and outliers, and subsequently performs poorly on unseen data. The optimal model (Figure 2.3b) achieves a balance between these two extremes, providing the most accurate predictions on unseen data [11], [13].

Techniques such as dropout, where random neurons are "dropped out" or turned off during training, can help mitigate overfitting. Another technique is data augmentation, which involves creating new training samples by applying transformations such as rotations, translations, and flips to the existing data. This helps prevent overfitting and allows the model to learn more robust features [11], [13]. The process of training deep learning models can demand alot of computational resources and may require the use of specific equipment, like GPUs/TPUs [11]. Moreover, the selection of hyperparameters, including the learning rate and the quantity of layers in the network, can notably impact the model's effectiveness. As a result, the process of training deep learning models frequently requires a trial and error approach, in order to discover the most efficient architecture and configurations [11].

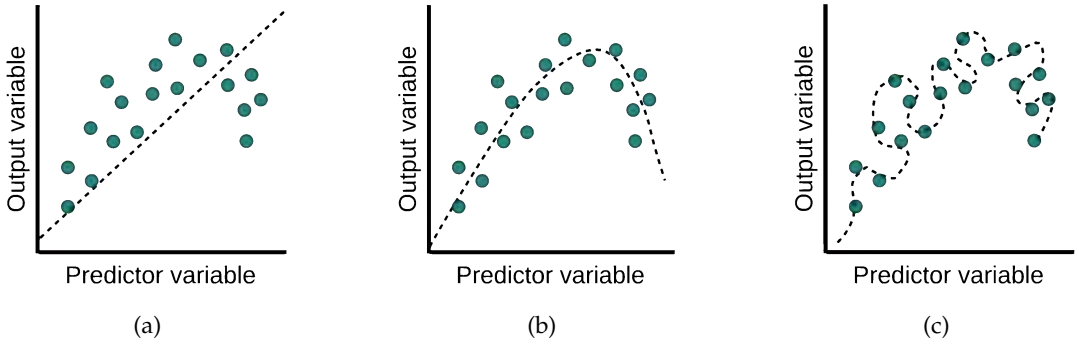


FIGURE 2.3: (a) underfit (b) optimal (c) overfit [15]

2.3 Image Segmentation

Image segmentation is a process in computer vision and image processing that partitions an image into multiple segments or sets of pixels, often referred to as superpixels [16]. The specific workings of image segmentation can vary greatly depending on the method used, but at a high level, it involves assigning labels to pixels to make it easy to differentiate between different regions of any image [17]. These could be color, intensity, texture, or other properties. The labeled pixels that share common characteristics form a segment. The primary goal of image segmentation is to simplify or change the representation of an image into something more meaningful and easier to analyze [17].

Building on recent survey papers and the continuous evolution of techniques in image segmentation, three methods have emerged as particularly impactful. These include Fully Convolutional Networks (FCNs), U-Net, and Mask R-CNN [17]. Each of these deep learning-based methods employs a neural network that is trained to recognize patterns in pixel values and their spatial relationships, leading to the assignment of each pixel to a particular class or segment [17].

Fully Convolutional Networks: The key advancement of FCNs is the transformation of fully connected layers into convolutional layers, which allows for input images of any size. Trained end-to-end, FCNs learn representations and produce dense pixel-wise predictions. They employ a method known as upsampling to map coarse predictions made by downsampled layers back to the original image size, thus accomplishing image segmentation [18].

U-Net: Distinguished by its U-shaped design, U-Net is a type of CNN that includes a contracting (downsampling) path to capture context and a symmetric expanding (upsampling) path for precise localization. U-Net shows exceptional efficiency in biomedical image segmentation tasks, where relevant features can be found at multiple scales within the image [19].

Mask R-CNN: Extending the popular object detection framework Faster R-CNN, Mask R-CNN adds a branch for predicting an object mask in parallel with the existing branch for bounding box recognition. This design enables Mask R-CNN to output accurate segmentations of objects in the image. The framework is flexible and can be implemented with a variety of backbone architectures [20].

The U-Net architecture (discussed in Chapter 2.4) and its variations have emerged as one of the most effective and widely used methods for cell segmentation [21], [22]. This approach has demonstrated superior performance compared to existing techniques, making it an ideal

choice for further exploration in the following chapters of this BA thesis [10], [22]. The U-Net model, with its unique design and underlying principles, excels at capturing intricate details and complex structures in cell images.

2.4 U-Net

The U-Net architecture is a deep-learning convolutional neural network; developed specifically for biomedical image segmentation, with the ability to produce highly detailed segmentation maps from a limited set of training samples. Introduced in 2015 by Ronneberger et al. [19], U-Net has since been widely adopted across various medical imaging applications due to its effectiveness and efficient use of labeled data.

The U-Net architecture is characterized by its U-shaped structure as shown in Figure 2.4, which consists of a contracting path to capture context and a symmetric expanding path that enables precise localization. The contracting path follows the typical architecture of a convolutional network, composed of repeated application of convolutions, each followed by a rectified linear unit (ReLU) and a max pooling operation. This path is responsible for capturing the context in the image. On the other hand, the expanding path consists of an upsampling of the features followed by a convolution, a concatenation with the correspondingly cropped feature map from the contracting path, and two regular convolutions. This path allows the network to use the context captured in the contracting path to localize and precisely segment the structures in the image [19].

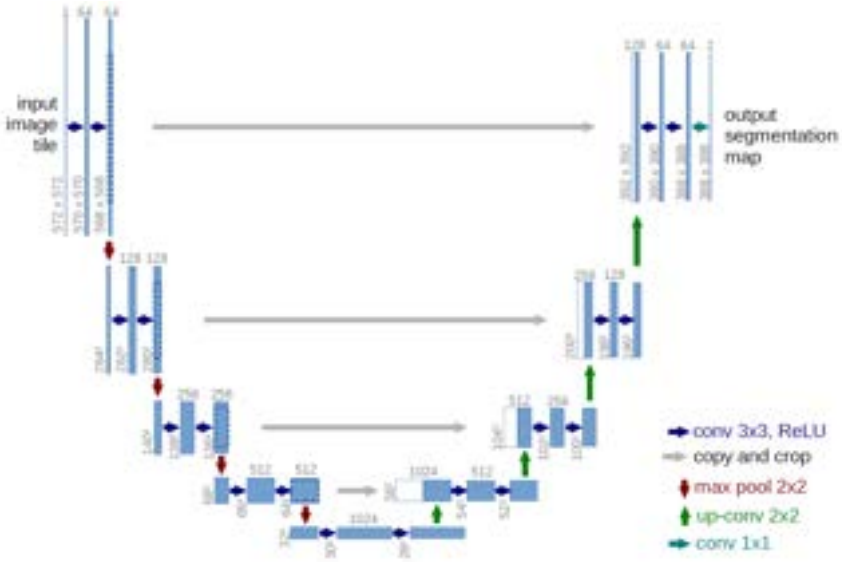


FIGURE 2.4: The original U-Net architecture [19]

In a comparative analysis of several cell segmentation methodologies, including U-Net, DeepCell, Random Forest, and CellProfiler, Caicedo et al. [10] demonstrated the superiority of deep learning strategies for segmenting nuclei's fluorescence images. Their study revealed that, amongst the evaluated approaches, U-Net excelled at segmenting individual cells and precisely outlining neighboring nuclei. It achieved this while avoiding common mistakes such as dividing single cells or merging multiple cells. U-Net surpassed the other models in all the evaluated metrics, including the highest Intersection over Union (IoU), lowest segmentation error rate, and the lowest rate of missed nuclei [10].

The architecture has been adapted for various tasks beyond the initial biomedical image segmentation. For instance, 3D U-net was developed to handle volumetric images, which is common in medical imaging. Attention U-net incorporated an attention mechanism into the U-net architecture to allow the model to focus on specific areas in the image. Inception U-net combined the Inception module with U-net to capture multi-level features in the image. These adaptations have been developed to address specific challenges in medical imaging and further demonstrate the flexibility and versatility of the U-net architecture [23].

Moreover, U-Net has been applied to a wide range of medical imaging modalities, demonstrating its versatility and robustness. In dermoscopy, U-Net has been used for skin cancer detection, specifically for melanoma, the most dangerous type of skin cancer [24]. In X-ray imaging, U-Net has been applied for the analysis of bones, including the diagnosis of rheumatoid arthritis and osteoporosis, and for the detection of pulmonary diseases in chest X-rays [25], [26].

U-Net's performance has also been demonstrated in the context of the recent COVID-19 pandemic. The medical imaging community has researched various deep-learning techniques, including U-Net, to diagnose COVID-19. The diagnostic images for COVID-19 are chest CT scan [27].

In Conclusion , the U-Net-based architecture has proven to be invaluable in the domain of medical image analysis. The surge in U-Net-related publications since 2017 reinforces its standing as a leading deep-learning approach in medical image diagnostics [23].

2.5 Leveraging Small Datasets in Deep Learning

The performance of a deep learning model can often be enhanced by adding more data to the training set [10], [28]. However, the availability of an abundance of labeled data, especially in specialized domains, is not always a given. This scarcity poses a challenge in training robust models, especially in fields such as medical imaging where procuring large volumes of annotated data can be labor-intensive, time-consuming, and subject to privacy concerns [29], [30]. The problem becomes even more pronounced when dealing with medical image data, given the high dimensionality and complexity of these datasets [31]. However, various research has proposed strategies to counteract the limitations imposed by small datasets. These strategies include data augmentation, transfer learning, and active learning [29], [31].

Data Augmentation

Data augmentation is a technique for artificially expanding the size of a dataset by creating transformed copies of images in the dataset. The transformations can include rotation, scaling, flipping, cropping, and color modification, among others. This technique can help improve the performance of a model by allowing it to learn from a greater variety of samples. Perez and Wang explored this strategy in their paper, "The Effectiveness of Data Augmentation in Image Classification using Deep Learning" [29]. Techniques suggested by [19], [29], [32] are:

- **Rotation:** Rotates the image by a specified angle.
- **Horizontal Flip:** Flips the image horizontally, i.e., about the vertical axis.
- **Vertical Flip:** Flips the image vertically, i.e., about the horizontal axis.
- **Scaling:** Changes the size of the image while maintaining the aspect ratio.

- **Shearing:** Distorts the image along an axis, creating a slanting effect.
- **Resizing:** Alters the resolution of the image, changing its dimensions.
- **Random Crop:** Extracts a smaller portion from the original image at a random location.

Transfer Learning

Transfer learning is a widely recognized machine learning strategy where knowledge gained from solving one problem is applied to solve related problems[33]. Essentially, a base network is first trained on a specific task and relevant dataset, then adapted to the target task with the relevant dataset [34].

The transfer learning process is mainly broken down into two steps: selecting a pre-trained model and determining the size and similarity of the problem [35]. The choice of pre-trained model depends on how closely the associated problem aligns with the target task. If the target dataset is small (e.g. less than 1000 images) and similar to the source training dataset (like medical, handwritten characters, vehicles, or biometric datasets), there's a higher risk of overfitting. Conversely, if the target data is large and similar to the source datasets, the risk of overfitting decreases, typically requiring only fine-tuning of the pre-trained model [34].

Transfer learning has demonstrated improved performance when the source and target tasks bear a closer resemblance [31]. Surprisingly, even transferring weights from significantly different tasks has shown better results than starting with random initialization [36]. Several studies have illustrated this. For instance, one research utilized weights from a general network (VGG16) and fine-tuned them for prenatal image segmentation in ultrasound imaging [37]. Another study applied original weights from a disparate application to polyp detection, requiring the fine-tuning of all layers [38]. This approach led to a 25% increase in sensitivity compared to only fine-tuning the last layer. Interestingly, some experiments reported superior results when training from scratch, as compared to fine-tuning a pre-trained network [39].

There are three primary levels at which transfer learning can be implemented [31]:

- **Full Network Adaptation**: This involves initializing weights using a pre-trained network (instead of a random initialization) and updating them all during training.
- **Partial Network Adaptation**: Network parameters are initialized from a pre-trained network, but the weights for the first few layers are frozen. Only the final layers are updated during training.
- **Zero Adaptation**: This strategy initializes the weights for the entire network from a pre-trained model without any changes.

In general, the zero adaptation approach is not recommended when transferring from another medical network due to significant variations in the appearance of the target organ. This is especially true if the source networks were trained on general images. Biomedical objects can greatly differ in appearance and size, meaning that transfer learning from models trained on images with significant cell variation may not improve segmentation results [31].

2.6 Dealing with Partially Labeled Data

As mentioned in Section 1.2, the process of manually labeling a single sample typically takes 10–15 minutes. Additionally, the dataset we obtained contains images that have been partially labeled. The annotators found it excessively time-consuming to label each image in its entirety. This situation highlights the importance of delving deeper into strategies for handling partially labeled data.

Weakly Supervised Learning

In their survey paper [40], Niclas Simmler, Pascal Sager et al. suggest weakly supervised approaches for partially labeled data. In general, weakly supervised approaches aim to extract high information predictions from labels that provide limited information [40].

One of the approaches mentioned [41] is an architecture to enhance the performance of a semantic segmentation network by leveraging information from various types of annotations, such as image-level labels, bounding-box labels, and pixel-level labels. The authors employ a fully convolutional network, a type of neural network commonly used for image analysis, to predict segmentation masks [41].

After obtaining the initial segmentation masks, they are passed through an annotation-specific loss module [41]. This loss module applies different loss functions based on the type of label available. By tailoring the loss function according to the label's form, the segmentation network can be improved effectively [41]. The results demonstrate that this approach effectively utilizes training data with different levels of supervision, indicating its ability to leverage various types of annotations to enhance the performance of the semantic segmentation network.

Semi-Supervised Self-Training

Another approach to address this challenge involves employing a semi-supervised learning paradigm that takes advantage of unlabeled data [42]. One particular technique within this paradigm is self-training [43]. Self-training is an iterative strategy that expands the labeled training sample set. It begins by initially training a model using labeled data. The model is then used to predict labels for the unlabeled data. Among the unlabeled data points, those with high confidence in their predicted labels are selected. These selected data points, along with their predicted labels (pseudo-labels), are gradually added to the training data [43].

In the research paper titled "Semi-Supervised Learning for Fine-Grained Classification With Self-Training" [44], the authors proposed a semi-supervised approach that utilizes self-training. By leveraging a supervised model to generate pseudo-labels for unlabeled data, the approach effectively enlarged the size of the training data. This enlargement contributed to improved performance in fine-grained classification tasks, as the model learned from both the initially labeled data and the newly incorporated pseudo-labeled data.

Cost-Effective Active Learning

Active learning is a technique where an algorithm can ask a human (or another source of information) to provide labels for specific pieces of data from a pool of unlabeled data [45]. The algorithm chooses which data to ask about based on how informative or uncertain the data is.

However, a method called Cost-Effective Active Learning (CEAL), proposed by Wang et al. in 2016 [46], takes a different approach. Instead of just focusing on the most informative data, it also considers unlabeled data. It does this by feeding unlabeled data into a type of algorithm called a CNN, and then choosing two types of data to fine-tune the CNN. The concept

of informativeness quantifies how much a single instance contributes towards reducing the uncertainty or imprecision in a statistical model [47].

One type of data it selects are those that the CNN isn't very confident about (low confidence level). These are considered informative because these instances are challenging for the model. By learning from these instances and adjusting its parameters to better predict them, the model can potentially improve its overall performance [46]. These pieces of data are then labeled by a human and added to the pool of labeled data.

The other type of data it selects are those that the CNN is very confident about (high confidence level). Because the CNN is so sure about these, the CEAL method gives them automatic labels, so no human needs to get involved. These two types of data - low and high confidence - help to balance each other out. This is because they represent different levels of confidence from the CNN about the unlabeled data. [46]

Several studies [45], [46], [48] implemented this approach and managed to leverage the majority pseudo-labeled samples to provide sufficient training data for robust feature learning.

3 Methods

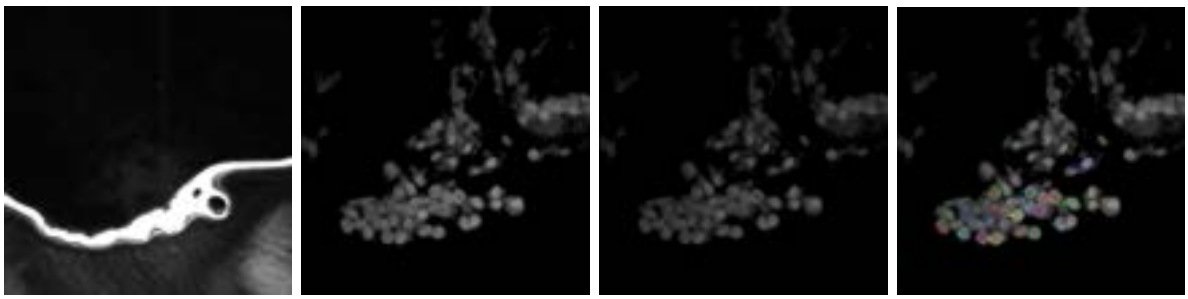
In this methods chapter, we present a concise overview of our methodology for cell segmentation on Fpol images. We cover the dataset used, preprocessing steps, transfer learning approach, modified U-Net architecture and post-processing techniques.

3.1 Dataset

We use a dataset provided by Prof. Dr. Anna Yaroslavsky's research team at the University of Massachusetts Lowell. The images, captured using a multi-channel confocal microscope, feature thyroid cells with illumination from two linearly polarized laser sources. Each image encompasses a field of view measuring $205, \mu m \times 205, \mu m$ and is recorded as 8-bit grayscale. The dataset includes reflectance, cross-polarized, co-polarized images, and annotated images, each with a resolution of 1000×1000 pixels in the TIFF file format [3].

The reflectance image (Figure 3.1a) visualizes the overall image intensity, while the co-polarized image (Figure 3.1b) captures the light that retains its polarization following interaction with the tissue. In contrast, the cross-polarized image (Figure 3.1c) displays light that has changed its polarization after tissue interaction. Lastly, the annotated image (Figure 3.1d) provides the ground truth for cell locations and boundaries.

It is essential to highlight the inherent subjectivity in the dataset annotation process, as seen in figure 3.1d. The researchers employ a checklist (see appendix A.1) to identify candidate cells for further examination. Notably, not all viable cells in a given image are typically annotated by researchers. Rather, they select a subset of the most suitable cells and proceed with subsequent analyses on this subset. As a result of this selective annotation, the ground truth data embodies an intrinsic ambiguity, as different researchers may favor slightly different cells for further exploration. This subjectivity introduces a form of incompleteness in the data, characterized by the potential absence of labels for certain viable cells.



(a) (b) (c) (d)

FIGURE 3.1: (a) Reflectance image (b) Co-polarized image (c) Cross-polarized image (d) Annotated image

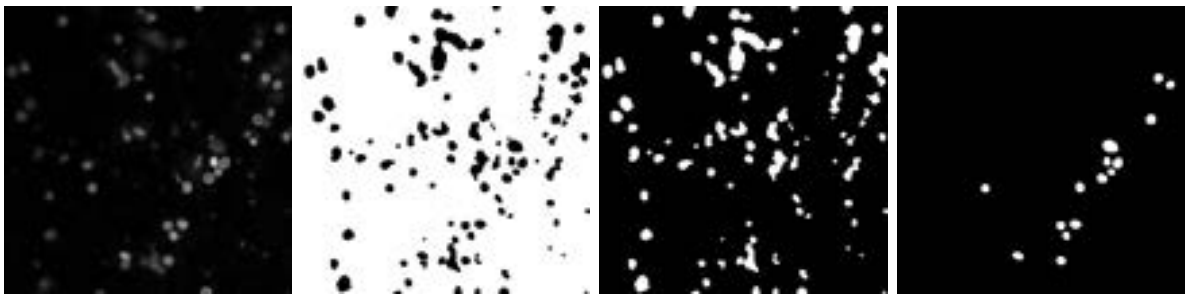
Preprocessing

The images from the dataset are processed to construct appropriate training and testing data for the cell segmentation task. The co-polarized images are selected as the input, as this is the image type primarily used by the researchers due to its higher contrast.

For the segmentation task, we established three classes: background, bad cells, and good cells. The good cells correspond to the cells annotated by the researchers. In contrast, the bad cells represent the remaining cells not segmented by the researchers. This decision to incorporate three classes, rather than the conventional binary classes (background and foreground) found in traditional cell segmentation tasks, is made to mitigate the dataset's inherent ambiguity. We hypothesized that the model would more easily distinguish between background and cells, thereby confining any potential confusion to the differentiation between good and bad cells.

The final dataset comprises 50 samples selected randomly from the initial collection. Each sample from the dataset includes a raw image (Figure 3.2a), a background mask (Figure 3.2b), and images of bad cells (Figure 3.2c) and good cells (Figure 3.2d) that should or should not be examined further, respectively.

We employed a pre-trained model to identify bad cells to perform an initial segmentation, followed by manual refinement to ensure accuracy. This pre-trained model is trained on the Kaggle Data Science Bowl 2018 dataset, which we will describe in a subsequent section.



(a) (b) (c) (d)

FIGURE 3.2: (a) Raw image (b) background mask (c) Bad cells (d) Good cells

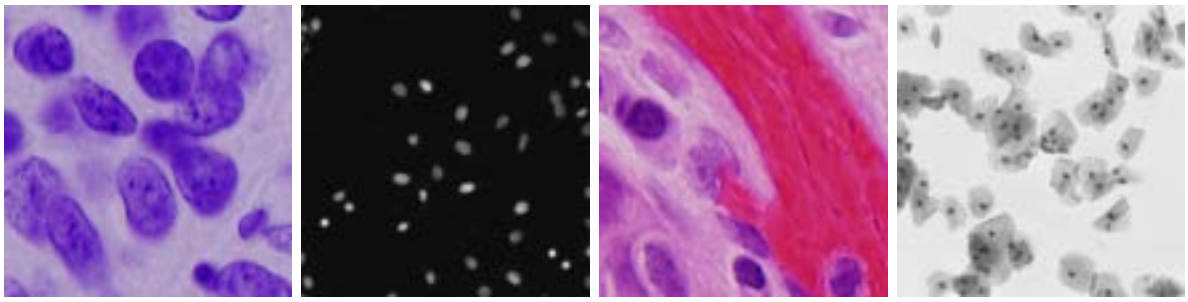
Transfer Learning

In addition to the primary dataset, the Data Science Bowl 2018 dataset, as described in Caicedo et al. [49], is utilized for transfer learning. This dataset was chosen due to its relevance to our task, as it is a large and diverse collection of annotated microscopy images of cell nuclei under various conditions. The conditions include purple tissue (Figure 3.3a), fluorescence (Figure 3.3b), combined pink and purple tissue (Figure 3.3c), and grayscale tissue (Figure 3.3d). The images were obtained from multiple sources, including samples, cell lines, microscopy instruments, imaging conditions, operators, research facilities, and staining protocols, making the dataset highly diverse and challenging for the models.

The dataset contains 670 images with a 256×256 or 320×256 pixels resolution. Each image is annotated with the nuclei's segmentation masks, clearly distinguishing between the background and cells.

This dataset was used for the 2018 Data Science Bowl competition, attracting 3,891 teams worldwide. The challenge was building a nucleus segmentation method that could be applied

to two-dimensional light microscopy image of stained cells. This aspect makes the dataset ideal for transfer learning in our current task.



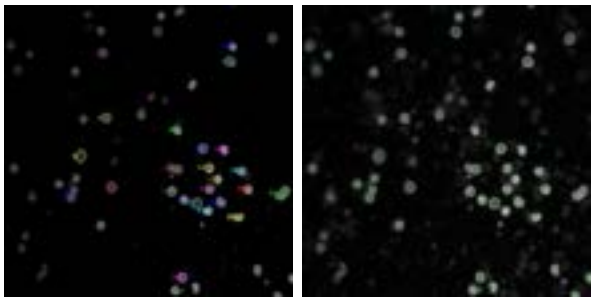
(a) (b) (c) (d)

FIGURE 3.3: (a) Purple image (b) fluorescence image (c) Purple and pink image (d) Grayscale tissue image [49]

Non-Ambiguous Data

The second experiment (refer to Section 4.3) is focused on examining the effects of utilizing non-ambiguous data. Each cell satisfying the predetermined research guidelines (see Appendix A.1) was annotated to eliminate any potential ambiguity within the dataset.

Figure 3.4 displays a representative sample of this non-ambiguous dataset. We can observe that in the image annotated by the researchers (Figure 3.4a), there exist cells that potentially warrant annotation. In contrast, the non-ambiguous sample (Figure 3.4b) exhibits a thorough annotation where every viable cell has been annotated.



(a) (b)

FIGURE 3.4: (a) base dataset sample (b) non-ambiguous sample

3.2 Model Architecture

The structure of the original U-Net architecture is preserved, with modifications made to adapt to the specific requirements of the dataset, as detailed in Section 3.1. Figure 3.5 illustrates the modified U-Net architecture; this adaptation incorporates an additional output class to categorize images into the “background”, “good”, and “bad” cells.

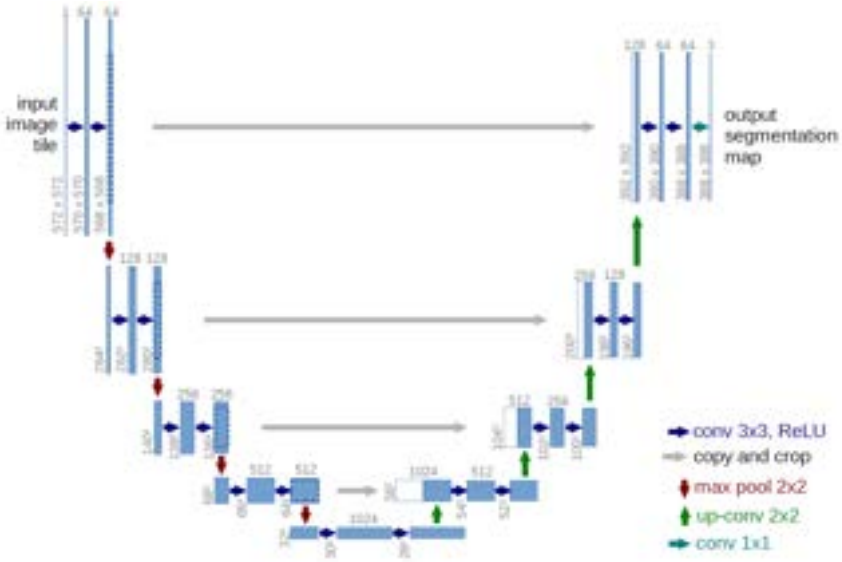


FIGURE 3.5: Modified U-Net architecture

In alignment with the transfer learning approaches discussed in the foundation section, this thesis introduces three distinct modifications to the U-Net architecture:

- **Implementation 1 - Original U-Net with Modifications:** This setup serves as the control model for comparison. It applies the modified U-Net design but without the use of any pre-training.
- **Implementation 2 - Full Network Adaptation:** This approach involves training the model on the transfer-learning dataset, then fine-tuning it on the target dataset.
- **Implementation 3 - Partial Network Adaptation:** This approach begins with the same pre-trained model from Implementation 2 but differs by freezing the encoding path.

Post-Processing

We employed two post-processing operations to enhance the quality of the segmented images and eliminate artifacts.

The first operation aims to rectify segmentation inconsistencies and enhance image clarity through a binary morphological process. We apply a dilation operation with a 3x3 kernel in three iterations, which expands segmented regions, thereby resolving small gaps in the objects. Following this, an erosion operation, mirroring the kernel size and iteration parameters of dilation, is performed to revert the dilated areas to their initial boundaries. This dual operation mitigates segmentation discontinuities, yielding a refined and seamless image representation.

The second operation is designed to eliminate potential artifacts in the segmented image. We implement an opening operation using an elliptical structuring element with a 15x15 kernel size, which offers versatility to accommodate diverse noise shapes. This operation involves a dilation-erosion sequence with the specified structuring element, substantially impacting isolated regions typified as noise or artifacts. Consequently, this operation significantly improves the segmentation quality and enhances the accuracy of the segmented objects.

4 Experiments & Results

This chapter delves into a detailed exploration of our experimental approach and the resulting outcomes. We initiate the process by establishing a baseline model, which relies solely on the initial dataset. The next phase seeks to improve this model's performance by incorporating non-ambiguous data. Further refining our methodology, we implement semi-supervised learning as well as active learning techniques to maximize the insights drawn from our limited image dataset. We conclude this chapter with a critical evaluation of our experimental design, outlining potential limitations and their implications.

4.1 Experimental Setup

The architecture used in this thesis is the modified U-Net structure outlined in section. The utilized loss function is cross-entropy loss, while the optimization algorithm is the adaptive moment estimation (ADAM) method. Additionally, several data augmentation techniques have been incorporated into the system using the Torchvision library [50]. These techniques include:

- Rotation
- Horizontal & Vertical Flip
- Scaling
- Shearing
- Resizing
- Random Crop

Quantitative Metrics

To evaluate the performance of the system quantitatively, several metrics were employed. These are chosen to ensure they measure key aspects of the model's output, particularly its ability to segment cells in images correctly. The metrics used are:

- **Intersection over Union (IoU)**: This metric provides a measure of overlap between the predicted segmentation and the ground truth. A higher IoU indicates a greater overlap and therefore better performance.
- **Sørensen-Dice Coefficient (dice)**: Similar to IoU, this coefficient also measures the similarity between the predicted and ground truth segmentations. It is beneficial in cases where the objects of interest are small or sparse within the image.
- **Precision**: This measures the accuracy of the positive predictions. Higher precision means that more predicted positives are true positives.
- **Recall**: Measures the ability of the system to identify all positive examples in the dataset. A higher recall indicates fewer false negatives.
- **F1 Score**: This is the harmonic mean of precision and recall, providing a single score that balances these two metrics. A higher F1 score indicates better overall performance.

Qualitative Metrics

In the qualitative evaluation, we assess the model's performance through a series of subjective metrics. These metrics address the model's ability to outline cell boundaries accurately, its performance in identifying and excluding cells that are blurry or out of focus, its ability to identify cells displaying high emission levels, its capacity to select cells that are not unusually small or large accurately, its capability to avoid merging distinct cells, and its restraint from unnecessarily splitting single cells.

We have further defined several qualitative metrics in collaboration with researchers from the University of Massachusetts. These metrics were selected based on their ability to provide comprehensive insights into different aspects of our model's performance. The following are the metrics used:

- **Total Cells Segmented by Model**: This is a quantitative measure that records the total number of cells our model was able to segment in the given dataset. This serves as a fundamental evaluation of the model's ability to identify and segment individual cells within the images.
- **Number of Correctly Segmented Cells**: This is the count of cells that the model segmented accurately, in alignment with the ground truth data. An increase in the number of correctly segmented cells is indicative of a high-performing model.
- **Number of Badly Segmented Cells**: This metric tracks the number of cells that were identified and segmented by the model but were later deleted because of errors or inaccuracies in the segmentation process. This measure helps us assess the precision of our model.
- **Precision**: This metric measures the accuracy of the model in segmenting cells. A high precision indicates that a large proportion of the cells segmented by the model are indeed correct.
- **False Discovery Rate (FDR)**: This metric provides a measure of the proportion of cells incorrectly identified by the model as compared to all the segmented cells. A lower FDR indicates that a lower proportion of the segmented cells were incorrect.

Expert Evaluation

Expert evaluation forms a vital part of the assessment process, providing first-hand feedback from researchers who interact with the model and evaluate its performance from a practical perspective. We provide the research team with an application programming interface (API) to facilitate seamless interaction with the model. Details regarding the usage of this API can be found in Appendix A.3.

One key factor is that it is easier for the researchers to delete incorrect cell identifications made by the model rather than manually segment the cells the model has missed. The reason for this is that segmentation is a time-consuming process, as described in Section 2.1.

The metrics for this part of the evaluation are:

- **Manual Segmentation Time**: This measures the time the research team takes to segment each sample manually.

- **Automatic Segmentation Time**: This is the time taken by the model to segment each sample automatically.
- **Manual Fpol Evaluation**: Florescence Polarization (Fpol) measure obtained from the manually segmented images.
- **Automatic Fpol Evaluation**: Florescence Polarization (Fpol) measure obtained from the models' segmented images.
- **Fpol Values Analysis**: This metric evaluates the effectiveness of the model in differentiating between benign and malignant cells based on Fpol values, which may be the most important metric for determining if the model can produce results comparable to expert evaluations.
- **Model Usefulness**: This is a subjective assessment from the research team about the overall utility of the model in their work. This qualitative metric is based on expert opinions and experiences while working with the model.

4.2 Experiment 1: Establishing a Baseline

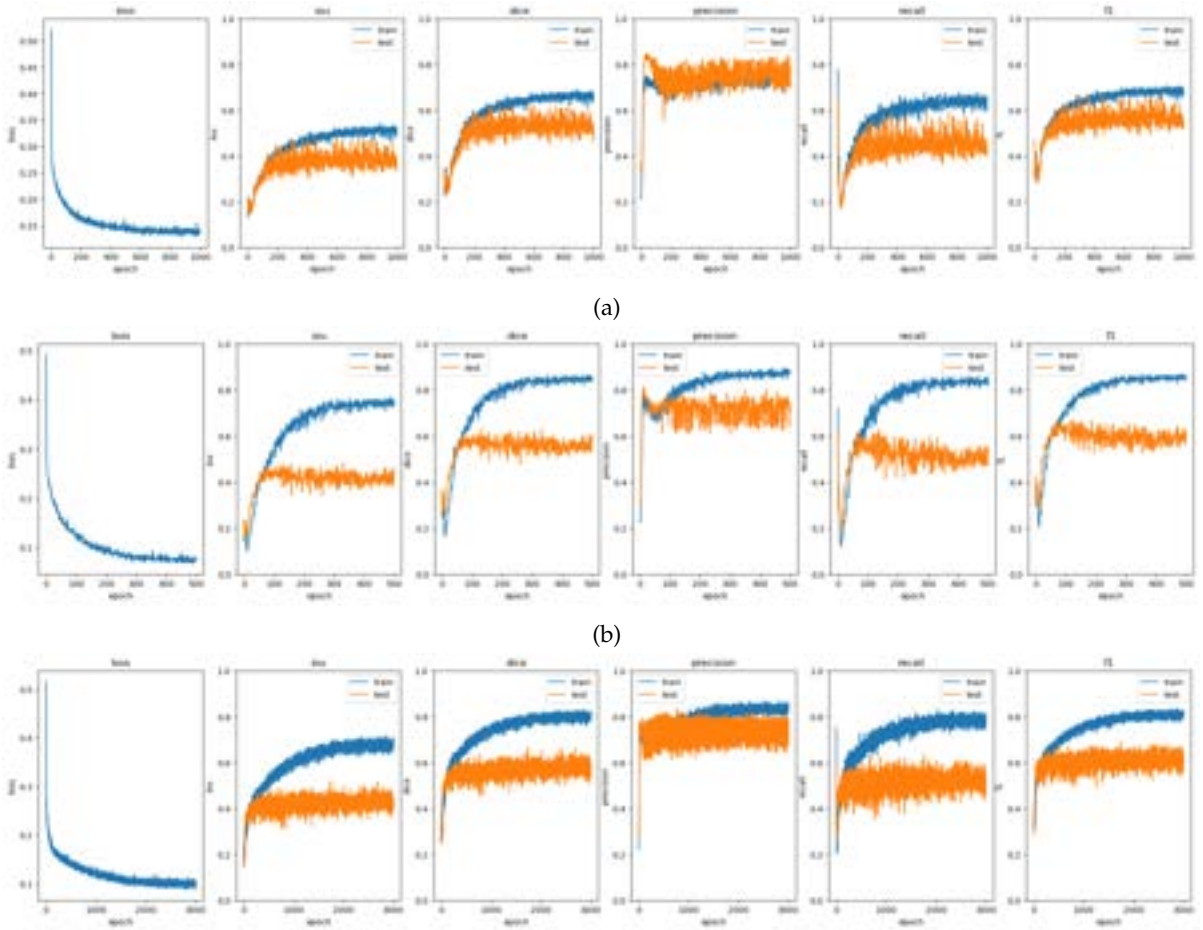
The primary objective of this experiment is to enhance the initial U-Net model, establishing a robust baseline for subsequent experiments. In order to achieve this, an experimental setup is constructed where we examine a selected range of hyperparameters identified through preliminary testing. Specifically, we opted for batch sizes of either 5 or 8, learning rates set at 1e-3 or 1e-5, and training epochs numbered at 500, 1000, or 3000. These hyperparameters are put to the test across three models: the modified U-Net, full network adaptation, and partial network adaptation. To determine the optimal set of parameters, a grid search is conducted within the defined hyperparameter space. This experiment operates on the original dataset (refer to Section 3.1), implementing a train/test split of 80/20 using seeded random sampling.

Quantitative Results

The results of the first experiment, which includes performance metrics from the baseline experiment for both training (tr) and testing (te) datasets, are listed in Table 4.1. These results were obtained by averaging the values from the final 30 epochs of each implementation, as depicted in Figure 4.1. The table details several metrics, including batch size (bs), learning rate (lr), epochs (ep), loss value, Intersection over Union (iou), Dice coefficient (dic), precision (pre), recall (rec), and F1-score (f1s). It compares the effectiveness of the three model implementations: the modified U-Net, full, and partial network adaptations.

Figure 4.1 provides a graphical representation of the performance evolution across epochs for the modified U-Net (Figure (a)), full network adaptation (Figure (b)), and partial network adaptation (Figure (c)) implementations. Each figure in Figure 4.1 presents the Intersection over Union (IoU), Dice coefficient, precision, recall, F1-score, and the loss function values, illustrating how these metrics evolved over the experiment's epochs. Detailed results for each implementation are provided in the Appendix section A.2.

- **Modified U-Net:** The Modified U-Net outperforms the other models on the testing set regarding IoU score, Dice coefficient score, precision, and F1-score. Despite subpar performance on the training set, compared to the full network adaptation, it demonstrates superior generalization when applied to the testing set.



(a)

(b)

(c)

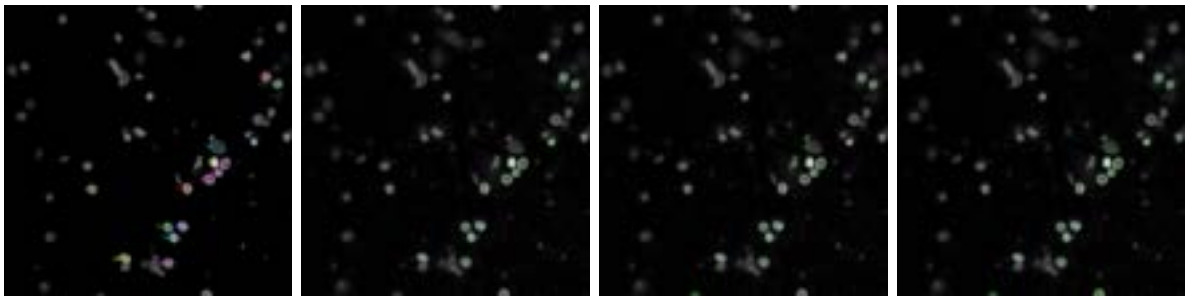
FIGURE 4.1: (a) modified network (b) full network adaptation (c) partial network adaptation

- **Full Network Adaptation:** Dominating on the training set, the Full Network Adaptation implementation achieves the lowest loss function value and secures the highest scores for IoU, Dice coefficient, precision, recall, and F1-score. Nonetheless, its performance lags on the testing set compared to the Modified U-Net.
- **Partial Network Adaptation:** Achieving the highest recall on the testing set, the Partial Network Adaptation is less effective in other testing and training metrics than the other two implementations. Figure 4.1c reveals its relatively lower values for training and testing sets, suggesting less effective learning.

Considering the testing dataset performance and generalization capability, the Modified U-Net implementation emerges as the best model among these three implementations.

TABLE 4.1: Baseline experiment quantitative metrics

net	bs	lr	ep	loss	tr-iou	tr-dic	tr-pre	tr-rec	tr-f1s	te-iou	te-dic	te-pre	te-rec	te-f1s
modified	8	1e-03	1000	0.105	0.666	0.792	0.820	0.781	0.800	**0.448**	**0.598**	**0.816**	0.500	**0.620**
full	5	1e-03	500	**0.075**	**0.742**	**0.846**	**0.873**	**0.833**	**0.853**	0.414	0.558	0.720	0.508	0.596
partial	5	1e-03	3000	0.102	0.678	0.801	0.837	0.782	0.808	0.426	0.576	0.754	**0.517**	0.613



(a) (b) (c) (d)

FIGURE 4.2: (a) annotated image (b) modified network (c) full network adaptation (d) partial network adaptation

Qualitative Results

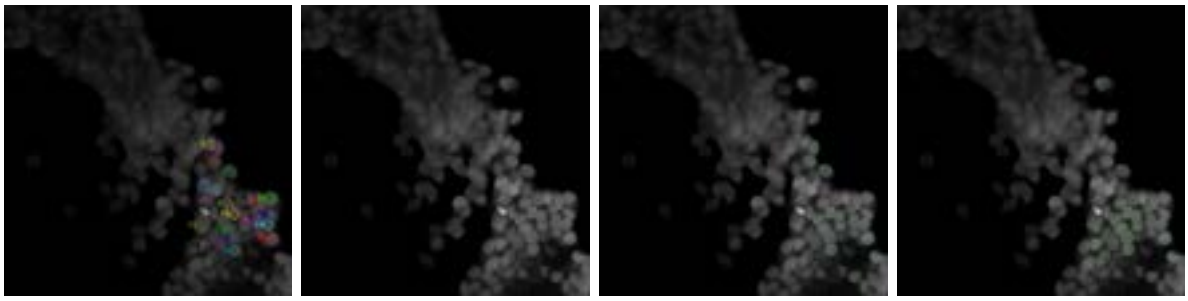
Figures 4.2 and 4.3 showcase a selection of images, illustrating the best and worst performances in terms of segmentation respectively. While the models generally identified cells, they often fell short of accurately determining cell boundaries, frequently leaving the segmentation incomplete or unclear. The full and partial adaptations generated numerous artifacts within the cells, as depicted in figures 4.3c and 4.3d.

- **Modified U-Net:** This model exhibited the lowest performance in the total number of segmented cells. An example of this can be seen in Figure 4.3b, where the model identified only two cells out of the 24 labeled by researchers. The model displayed a marked preference for cells with high fluorescence intensity.

- **Full Network Adaptation:** This adaptation produced comprehensive cell segmentation results. Compared to the modified model, it tended to limit cells more precisely up to their boundaries. The increased cell segmentation is likely attributable to the prior learning on the Data Science Bowl dataset discussed in Section 3.1.

- **Partial Network Adaptation:** This model segmented the most cells overall, though the quality of the segmentation was comparatively lower. It often failed to clearly distinguish the boundaries of the cells, frequently clustering groups of cells into single segmentations and selecting blurry or out-of-focus cells. We hypothesize that this is a consequence of the frozen encoding path, which limits the model's ability to adapt to the new task.

The Full Network Adaptation demonstrated the best performance in this baseline assessment. It segmented more cells than the modified model, maintaining a relatively high segmentation quality without unnecessary cell merging or splitting. Table 4.2 displays the performance of the Full Network Adaptation model on ten samples (samples listed in Appendix section A.4). The model averaged a precision of 77.4%, indicating the correctly segmented cells, but missed about 3.9 cells per sample. It also had a False Discovery Rate of 18.3%, pointing to the proportion of incorrect cell predictions. From these results, it can be seen that while the Full Network Adaptation model exhibits decent performance in terms of precision (77.4%), there are areas for improvement, especially in reducing the number of cells deleted and undetected.



(a) (b) (c) (d)

FIGURE 4.3: (a) annotated image (b) modified network (c) full network adaptation (d) partial network adaptation

TABLE 4.2: full network adaptation qualitative results

Sample	Cells segmented	Cells deleted	Cells not detected	Precision	False Discovery Rate
1	43	12	6	0.721	0.162
2	32	9	6	0.719	0.207
3	3	1	3	0.667	0.600
4	55	10	5	0.818	0.100
5	33	4	7	0.879	0.194
6	41	2	2	0.951	0.049
7	24	5	4	0.792	0.174
8	30	6	3	0.800	0.111
9	18	4	2	0.778	0.125
10	13	5	1	0.615	0.111
avg	29.2	5.8	3.9	0.774	0.183

4.3 Experiment 2: Impact of Non-Ambiguous Data

The second experiment assesses the potential benefits of introducing non-ambiguous data to the baseline model. The core objective is to evaluate whether such data could increase the model's performance. For this, we fine-tune all models defined in Experiment 1 using a new training set of 16 images in subsets of 2, 4, 8, and 16 while maintaining the same testing set as the previous experiment.

Results

The results of the second experiment, which includes performance metrics from the non-ambiguous data experiment for both training (tr) and testing (te) datasets, are summarized in Table 4.3, where (bs), (lr), and (ep) represent the batch size, learning rate, and epochs, respectively, identifies the leading model implementations. It details several metrics, including the loss value, Intersection over Union (IoU), Dice coefficient (dic), precision (pre), recall (rec), and F1-score (f1s).

- **Modified U-Net:** This implementation exhibits superior performance across all metrics on the training and testing sets, demonstrating the best IoU score, Dice coefficient score, precision, recall, F1-score, and the lowest loss function value. This suggests that introducing non-ambiguous data has considerably enhanced the modified U-Net model.

TABLE 4.3: Non-Ambiguous data experiment quantitative metrics

net	bs	lr	ep	loss	tr-iou	tr-dic	tr-pre	tr-rec	tr-f1s	te-iou	te-dic	te-pre	te-rec	te-f1s
modified	8	1e-03	3000	**0.025**	**0.900**	**0.947**	**0.952**	**0.943**	**0.947**	**0.574**	**0.722**	0.637	**0.876**	**0.738**
full	5	1e-03	3000	0.032	0.874	0.933	0.938	0.928	0.933	0.562	0.706	**0.643**	0.831	0.725
partial	8	1e-03	3000	0.067	0.819	0.900	0.907	0.894	0.900	0.518	0.676	0.595	0.823	0.690

- **Full Network Adaptation:** This model yields the second-best performance metrics. While its scores are lower than those of the modified U-Net, the full network adaptation model exhibits a higher precision on the testing set. However, its lower recall offset this advantage, resulting in a slightly lower overall F1-score.

- **Partial Network Adaptation:** This implementation scores lowest on all testing and training metrics among the three models.

Figure 4.4 presents the performance evolution across epochs for the modified U-Net (Figure (a)), full network adaptation (Figure 4.4b), and partial network adaptation (Figure 4.4c) implementations. Each figure depicts the Intersection over Union (IoU), Dice coefficient, precision, recall, F1-score, and the loss function values. As the figure demonstrates, all models show reduced variability along the y-axis compared to the previous experiment, indicating a more consistent performance across different metrics. Additionally, the performance of the models remains consistently stable without any noticeable trend of improvement or decline, even after extended fine-tuning.

Figure 4.5 presents the performance evolution across different sample sizes for the modified U-Net, showcasing the Intersection over Union (IoU), Dice coefficient, precision, recall, and F1-score. Despite the lower precision noted when training with non-ambiguous data, a significant increase in recall is observed, underlining the overall benefits of such data. On the other hand, IoU and Dice coefficient see a slight improvement, while precision experiences a noticeable decrease. This decrease in precision can be attributed to the ambiguity in the testing dataset.

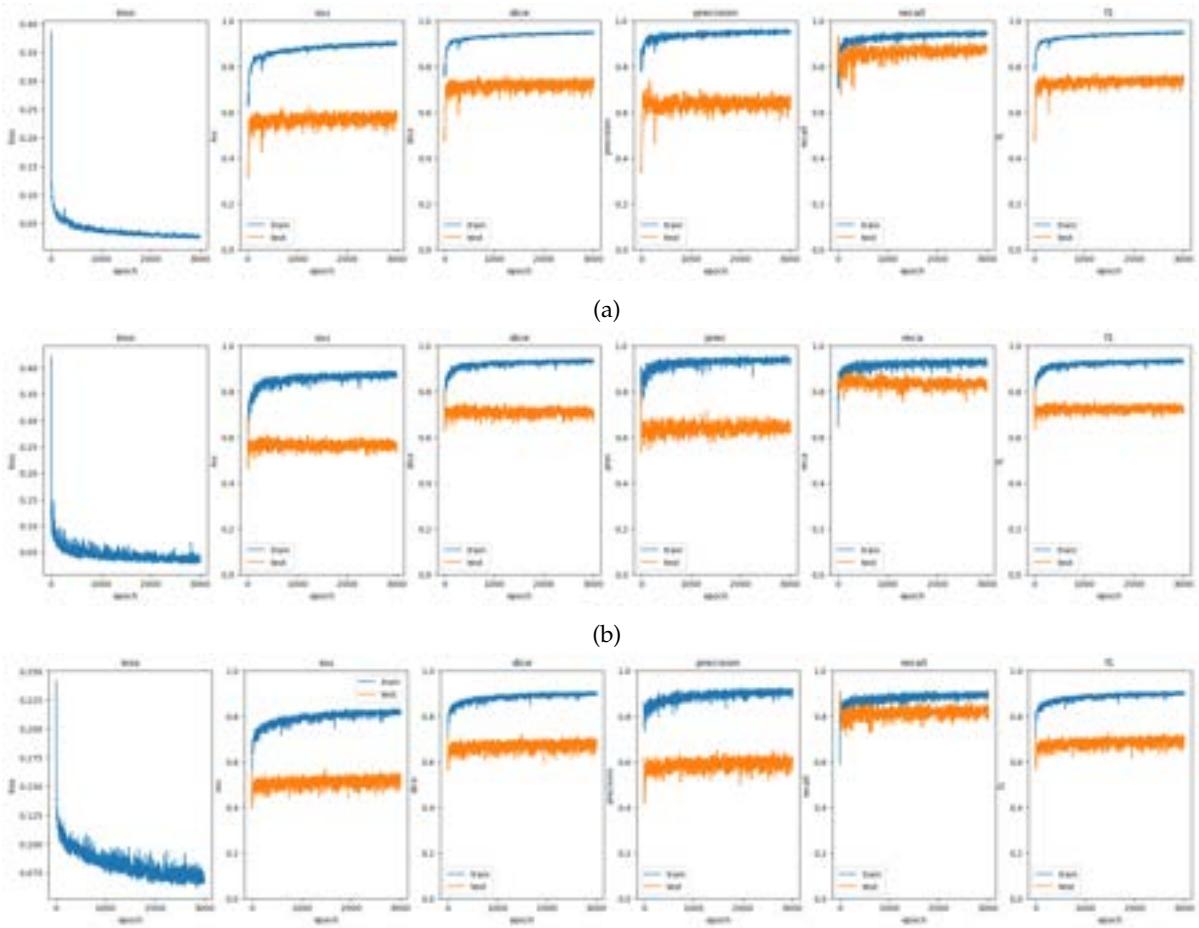
Given these insights, the Modified U-Net again emerges as the most effective model, exhibiting remarkable generalization ability and superior performance across various metrics.

Qualitative Results

Figures 4.6 and 4.7 show images; mirroring the ones used in the previous experiment. In this comparison, we observe that all models have improved the quality of segmented cells, as well as the number of cells they segmented. Both the full and partial adaptations, as depicted in figures 4.7d and 4.7c, demonstrated a decrease in artifacts when compared to the previous experiment.

- **Modified U-Net:** This model recorded the most significant increase in the number of segmented cells. Despite this improvement, the quantity still fell short of the other two implementations. Notably, the introduction of non-ambiguous data had the most considerable impact on this model.

- **Full Network Adaptation:** This model, compared to the baseline implementation, showcased improved specificity, illustrating its superior capability to distinguish between "good" and "bad" cells. This means that it effectively excluded blurry and out-of-focus cells. Additionally, it segmented more cells with high emission, evident in Figure 4.6c where the full adaptation segmented all high emission cells.



(a)

(b)

(c)

FIGURE 4.4: (a) modified network (b) full network adaptation (c) partial network adaptation

- **Partial Network Adaptation:** This model enhanced its ability to segment cells up to the cell boundary correctly. The most significant improvement was that it lowered its tendency to cluster groups of cells together, a change that becomes apparent when comparing Figures 4.3d and 4.7d.

The Full Network Adaptation stood out with the best performance. It produced better cell segmentations while offering more precise segmentations compared to the other two models. The model's advanced performance is quantitatively reinforced by the data presented in Table 4.4, detailing the results of the non-ambiguous data experiment.

Comparing the average results from the first experiment (exp1) and the second experiment (exp2), the Full Network Adaptation demonstrated marked improvements across all the metrics. Specifically, the number of cells segmented rose from an average of 29.2 in exp1 to 42.3 in exp2. At the same time, the number of deleted cells decreased from 5.8 to 3.9 on average, and the count of undetected cells fell from 3.9 to 1.9 on average. These advancements led to an improvement in precision from 77.4% to 89.0%. Meanwhile, the False Discovery Rate saw a significant reduction from 18.3% in exp1 to 6.3% in exp2, indicating a decrease in incorrect cell predictions.

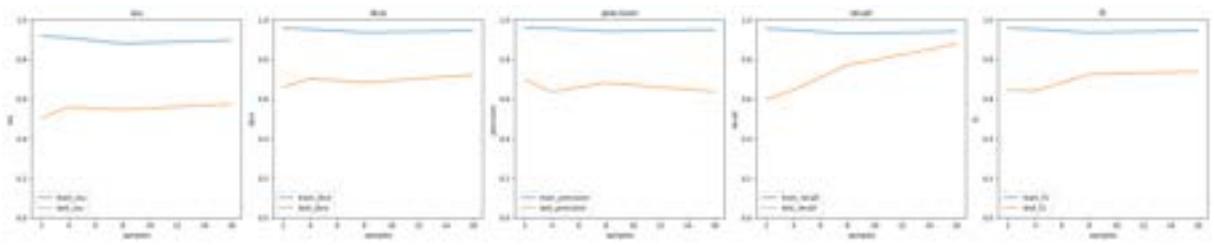
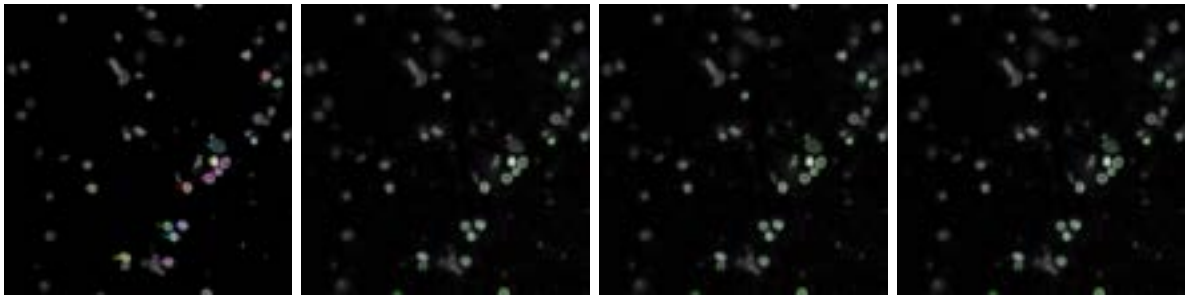


FIGURE 4.5: Non-Ambiguous Data samples evolution



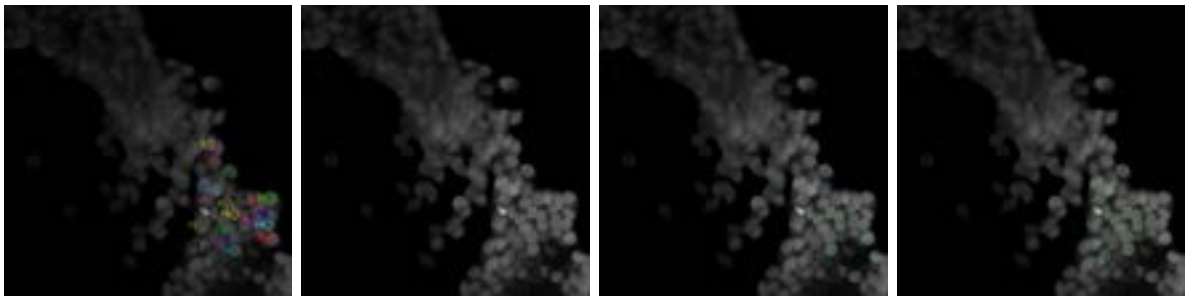
(a) (b) (c) (d)

FIGURE 4.6: (a) annotated image (b) modified network (c) full network adaptation (d) partial network adaptation

4.4 Experiment 3: Semi-Supervised Active Learning

To reduce the time taken for manual annotations, we adopted a strategy involving semi-supervised learning (SSL) and semi-supervised active learning (SSAL), leveraging pseudo-labeled images. However, unlike the models discussed in Chapter 2.6 that use a confidence-based approach to decide on the unlabeled data for pseudo-labeling, we integrated a manual selection process by human annotators. Our decision was motivated by the restricted size of our total sample base, comprising roughly 90 images, the majority of which had been used for Experiment 1 (Chapter 4.2) and Experiment 2 (Chapter 4.3).

For this third experiment, we again chose the best-performing model from Experiment 2, as



(a) (b) (c) (d)

FIGURE 4.7: (a) annotated image (b) modified network (c) full network adaptation (d) partial network adaptation

TABLE 4.4: full network adaptation qualitative results

Sample	Cells segmented	Cells deleted	Cells not detected	Precision	False Discovery Rate
1	64	6	3	0.906	0.049
2	57	8	3	0.860	0.058
3	9	3	3	0.667	0.333
4	77	4	4	0.948	0.052
5	54	6	3	0.889	0.059
6	44	1	1	0.977	0.023
7	45	1	1	0.978	0.022
8	40	8	1	0.800	0.030
9	19	1	0	0.947	0.000
10	14	1	0	0.929	0.000
avg exp1	29.2	5.8	3.9	0.774	0.183
avg exp2	42.3	3.9	1.9	0.890	0.063

determined by quantitative and qualitative metrics. We utilized this model to segment 14 manually selected unlabeled images, focusing on images where the cancer cells were not tightly clustered, exhibited distinct cell boundaries, and showed no blurriness. This targeted selection was primarily aimed at mitigating weaknesses identified in the model from Experiment 2, thus increasing its segmentation performance.

Our test dataset remains the same as in Experiments 1 and 2. The segmentations produced by this procedure serve as new ground truth data, feeding into both the SSL and SSAL strategies. In the SSL method, these segmentations are directly incorporated. In contrast, the SSAL approach incorporates an additional step wherein a human annotator refines the images before they are deemed suitable as ground truths.

Results

The results of the third experiment, including the performance metrics from the SSL and SSAL experiment for both training (tr) and testing (te) datasets, are summarized in Table 4.5. As in the previous experiments, (bs), (lr), and (ep) denote the batch size, learning rate, and epochs, respectively, with metrics including the loss value, Intersection over Union (iou), Dice coefficient (dic), precision (pre), recall (rec), and F1-score (f1s).

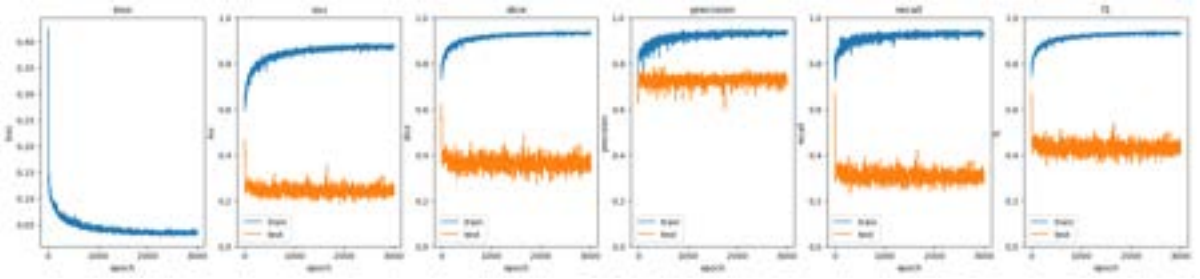
The SSL and SSAL models rapidly converge to a stable level. For the SSL model, the test dataset metrics instantly become smaller and remain stable, while the train set metrics stabilize after about 1,500 epochs. On the other hand, the SSAL model's metrics decrease gradually and do not reach as low a level as the SSL model, achieving a stable level in both the test set and the training set after approximately 1,000 epochs. Notably, the precision in the test set remains almost constant for both models.

- **Semi-Supervised Learning:** This approach shows superior performance in the training set across all metrics, demonstrating the best IoU score, Dice coefficient score, precision, recall, F1-score, and the lowest loss function value. However, its performance in the test set is less impressive, particularly with lower IoU and Dice coefficient scores, though it does have the highest precision.

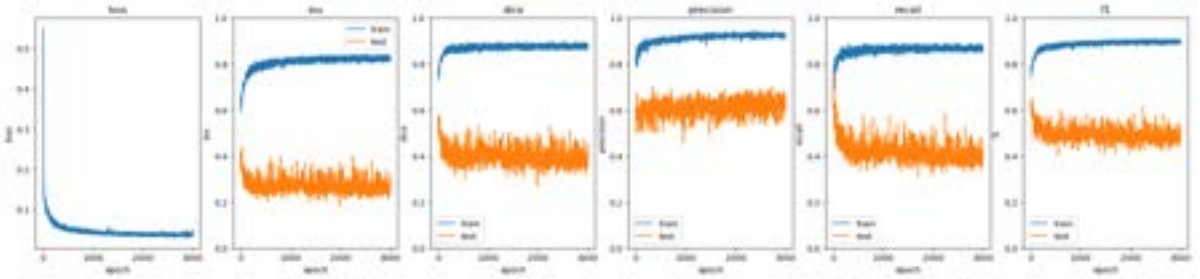
- **Semi-Supervised Active Learning:** While this method's metrics on the training set are

TABLE 4.5: Semi-supervised learning quantitative results

net	bs	lr	ep	loss	tr-iou	tr-dic	tr-pre	tr-rec	tr-f1s	te-iou	te-dic	te-pre	te-rec	te-f1s
SSL	8	1e-03	3000	**0.038**	**0.873**	**0.932**	**0.935**	**0.929**	**0.932**	0.245	0.364	**0.734**	0.307	0.432
SSAL	8	1e-03	3000	0.041	0.825	0.878	0.926	0.870	0.897	**0.262**	**0.385**	0.623	**0.395**	**0.483**



(a)



(b)

FIGURE 4.8: (a) Semi-Supervised Learning (b) Semi-Supervised Active Learning

not as high as SSL, it outperforms SSL on the testing set for IoU, Dice coefficient, recall, and F1-score. This suggests that SSAL might generalize better to unseen data.

Figure 4.8 presents the performance evolution across epochs for the SSL (Figure 4.8a) and SSAL (Figure 4.8b) strategies. Each figure illustrates the Intersection over Union (IoU), Dice coefficient, precision, recall, F1-score, and the loss function values. As shown, both SSL and SSAL quickly converge to a stable level.

When comparing SSAL to SSL, SSAL demonstrated better generalization to unseen data, indicating its potential advantage in real-world applications. However, when compared with the results from Experiment 2, the introduction of both SSL and SSAL resulted in a performance decline. Possible explanations for this behaviour are discussed in the Qualitative section.

Qualitative Results

The outcomes of Experiment 3, as detailed in Table 4.6, demonstrate the efficacy of the semi-supervised learning (SSL) approach in enhancing data production, with an average completion time for automatic segmentation of just 14.3 seconds, a substantial improvement on the timescales reported in Experiment 2 (Section4.3).

However, while faster, SSL does not match the semi-supervised active learning (SSAL) method in terms of precision or false discovery rate. Tables 4.6 and 4.7 contrast these performance metrics, showing the superior precision of the SSAL method.

TABLE 4.6: Semi-Supervised Learning qualitative results

Sample	Cells segmented	Cells deleted	Cells not detected	Automatic segmentation time [sec]	Precision	False Discovery Rate
1	33	7	15	17	0.788	0.366
2	27	11	11	17	0.593	0.407
3	9	3	2	9	0.667	0.250
4	40	12	12	15	0.700	0.300
5	26	10	13	14	0.615	0.448
6	43	4	4	16	0.907	0.093
7	32	2	4	15	0.938	0.118
8	30	3	3	10	0.900	0.100
9	20	2	2	15	0.900	0.100
10	7	1	7	15	0.857	0.538
avg	26.7	5.5	7.3	14.3	0.786	0.272

TABLE 4.7: Semi-supervised active learning qualitative results

Sample	Cells segmented	Cells deleted	Cells not detected	Manual annotation time [min]	Assisted annotation time [min]	Precision	False Discovery Rate
1	62	5	2	25	10	0.919	0.034
2	43	6	3	19	8	0.860	0.075
3	10	1	1	7	2	0.900	0.100
4	63	5	4	24	9	0.921	0.065
5	50	4	3	19	10	0.920	0.061
6	47	0	0	20	3	1.000	0.000
7	53	0	0	22	3	1.000	0.000
8	44	4	2	18	8	0.909	0.048
9	24	0	0	16	5	1.000	0.000
10	10	1	1	8	4	0.900	0.100
avg	40.6	2.6	1.6	17.8	**6.2**	0.933	0.048

Moreover, the SSAL approach increases efficiency, reducing annotation time by roughly 65% compared to Experiment 2 (Section 4.3), without compromising result quality. Table 4.8 verifies this, with SSAL outperforming SSL in precision and false discovery rate. While SSL offers speed, SSAL presents an optimal blend of efficiency and precision, making it more suitable for cell segmentation tasks.

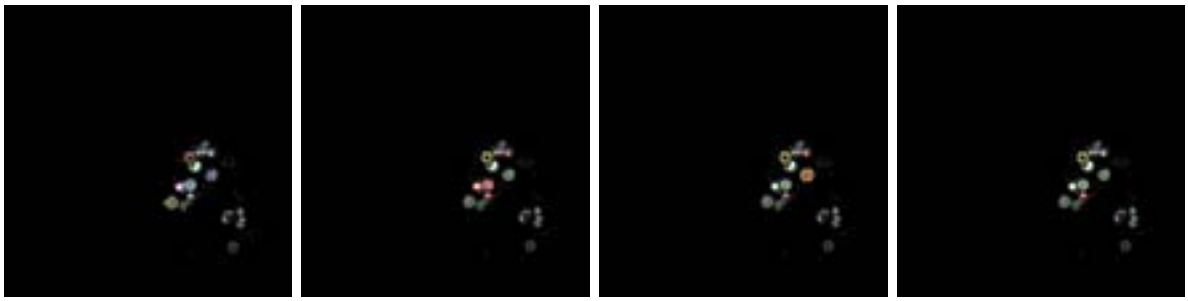
These outcomes also find visual representation in Figure 4.9, demonstrating cell segmentation under various experimental conditions. In the figure, red-marked cells indicate incorrect segmentation, yellow-marked cells were overlooked by the model but selected by researchers, and the correctly segmented cells are outlined in green.

The less optimal segmentation from Experiment 2 (Figure 4.9b) is evident compared to SSL and SSAL. Experiment 2 often clustered distinct cells into one, resulting in useless segmentation. Even though SSL (Figure 4.9c) shows an improvement, it struggles with full-cell segmentation. In contrast, SSAL (Figure 4.9d) delivers the highest overall performance, with most cells correctly identified and segmented.

In summary, while SSL brings speed benefits to cell segmentation, the SSAL method ensures

TABLE 4.8: Qualitative results average comparison

Sample	Cells segmented	Cells deleted	Cells not detected	Precision	False Discovery Rate
avg exp1	29.2	5.8	3.9	0.774	0.183
avg exp2	42.3	3.9	1.9	0.890	0.063
avg SSL	26.7	5.5	7.3	0.786	0.272
avg SSAL	40.6	2.6	1.6	0.933	0.048



(a) (b) (c) (d)

FIGURE 4.9: (a) full adaptation exp2 (b) full adaptation exp3 (c) full adaptation SSL (d) full adaptation SSAL

superior precision and a lower false discovery rate. These findings underline a potential trade-off between speed and quality in cell segmentation, emphasizing the importance of the method selection according to specific experimental or operational needs.

The discrepancy between quantitative and qualitative performance of the models from Experiment 2 and 3 may be due to factors such as variation in dataset composition, enhanced generalization from the SSAL model, potential model overfitting, and the nature of performance metrics used. The integration of 14 additional non-ambiguously labeled images could have introduced variety, improving generalization but not necessarily quantitative metrics. Similarly, overfitting to unique features in the SSAL model might enhance qualitative performance without benefiting quantitative measures. This highlights the importance of combining both evaluation methods for comprehensive performance assessment.

Expert Evaluation

The research team noted substantial reductions in the time required to evaluate a sample, citing the model as particularly beneficial. According to feedback, the model used in Experiment 2 (Section 4.3) occasionally grouped distinct cells, but despite this shortcoming, its overall usefulness in expediting the segmentation process was highly appreciated.

To assess the practical applicability of our model's segmentations, it is crucial to ensure that the Fpol values derived from these segmentations closely match those obtained from manual segmentations conducted by domain experts. The evaluation of Fpol measures are currently in progress and remains to be completed at the time of this writing. These measures require significant computation and evaluation time. Consequently, the results of Fpol measures will be included in an updated version of this thesis.

Table 4.9 showcases the comparative times for manual and automatic segmentation and a qualitative assessment of the model's usefulness. Each sample took 1-2 hours to segment manually,

TABLE 4.9: Semi-Supervised Learning qualitative results

Sample	Manual segmentation time [h]	Automatic segmentation time [sec]	Fpol manual	Fpol automatic	Fpol evaluation	Model usefulness
1	1-2	20	tbd	tbd	tbd	very useful
2	1-2	17	tbd	tbd	tbd	very useful
3	1-2	8	tbd	tbd	tbd	very useful
4	1-2	15	tbd	tbd	tbd	very useful
5	1-2	15	tbd	tbd	tbd	very useful
6	1-2	15	tbd	tbd	tbd	very useful
7	1-2	15	tbd	tbd	tbd	very useful
8	1-2	8	tbd	tbd	tbd	very useful
9	1-2	15	tbd	tbd	tbd	very useful
10	1-2	15	tbd	tbd	tbd	very useful

while automatic segmentation ranged from 8 to 20 seconds, which results in a time reduction of 99.7%. Despite the lack of data for the Fpol measures (tbd - to be determined), the model's utility was consistently rated as "very useful" across all ten samples.

This feedback from the research team underscores the model's ability to significantly decrease segmentation time, a benefit that renders the model a valuable tool in this context. While there is room for improvement, such as occasional cell clustering and empty-space segmentation, it upholds the overall utility of the model. The latest SSAL model developed in Experiment 3 offers potential solutions to these identified limitations.

4.5 Limitations

Despite the promising results obtained from our experiments, there are several limitations that must be acknowledged in this study:

Limited Sample Pool: One of the most salient limitations lies in our restricted dataset. With only 90 images available for training and testing, our model had a relatively narrow scope of data to learn from. This constraint might impact the generalizability of the model, limiting its performance on new, unseen data [51].

Cell Clustering: Our model encountered difficulties when images contained high-density cell clusters (see Figure 4.7). Due to the limited training data available, the model's performance was less satisfactory when dealing with tightly packed cell populations. Segmentations were often smaller than the cell outline.

Microscope Variability: The model may struggle to generalize to images obtained from different microscopes and machines, as these can produce images with significant differences in aspects such as contrast and brightness. Without additional training on such varied data, the model's performance could be compromised under these conditions.

Generalizability to Other Cancer Types: While the proposed method focuses on thyroid cancer cell segmentation, its generalizability to other types of cancer may be limited. Each cancer type exhibits unique characteristics and imaging features, which may require tailored approaches or additional data for accurate segmentation. Evaluating the model's performance on diverse cancer types would provide insights into its broader applicability.

Need for More Non-Ambiguous Data: In the second experiment, we used only 16 non-ambiguously labeled images for fine-tuning. More rigorous testing is required to ascertain if there is a point of saturation beyond which the inclusion of additional non-ambiguous data ceases to yield substantial improvement in the model's performance.

Reliance on Manual Selection for Pseudo-Labeling: In our approach, we opted for human annotators to select the images to be pseudo-labeled, which inherently introduces a degree of subjectivity and potential bias in the selection process. An automated selection method, albeit more challenging to design, could possibly mitigate this limitation.

Dependence on Quality of Manual Annotations: The success of our approach largely hinges on the quality of manual annotations which were used to train the model initially. Any inconsistencies or inaccuracies in these initial labels could propagate through to the model's learning process and subsequently affect its performance.

Scalability and Deployment: While our approach showed promising results, considerations should be given to the scalability and deployment of the model in real-world clinical settings. Our model does not automate the entire process of thyroid cancer diagnosis. Post-segmentation, it is necessary to manually load the segmented images into another software or platform for further analysis to calculate fluorescence polarization values, which is another critical step in the diagnosis process. The absence of a comprehensive, end-to-end diagnostic solution is an area of limitation that future research could address. Factors such as computational requirements, integration with existing healthcare systems, and the potential need for further optimization should also be addressed to facilitate the practical implementation of the proposed method.

Lack of Real-time Application: The proposed approach may not be suitable for real-time or near real-time applications due to the computational complexity and processing time required for cell segmentation. The delay between image acquisition and obtaining segmentation results could limit its practical utility in time-sensitive clinical scenarios.

Computational Resource Requirements: Deep learning models, particularly those with complex architectures like U-Net, can be computationally intensive and require substantial computational resources for training and inference. The high computational requirements may pose limitations in terms of time, hardware, or accessibility, hindering the widespread adoption and deployment of the proposed method in resource-constrained settings.

Lack of Expert Evaluation for Fpol Calculation: An important limitation of this study is the absence of expert evaluation on the accuracy of the Fpol values derived from the segmented images produced by our model. The calculation of Fpol values is a crucial step in diagnosing thyroid cancer, and while our model aids in efficient segmentation, its effectiveness in producing images that result in accurate Fpol calculations has not been rigorously assessed by domain experts.

5 Conclusions

In this final chapter, we bring our bachelor's thesis to a close by presenting a concise summary of the key findings and insights garnered from our research. This summary encapsulates the main contributions of our thesis, highlighting the novel aspects and significant results obtained during the experimentation and evaluation phases. Additionally, we outline potential avenues for future studies, highlighting the need for further refinement and exploration in the field of deep learning assisted thyroid cell segmentation.

5.1 Summary & Discussion

This thesis implements and evaluates an automated cell segmentation method for the diagnosis of thyroid cancer by introducing a modified U-Net architecture for cell segmentation in Fpol images.

The baseline model, as assessed in Experiment 1, encountered challenges when dealing with clustered or overlapping cells, underscoring a requirement for more refined techniques. The partially adapted network, however, despite performing worse overall, displayed an improved capability in segmenting such complex cell formations, indicating the potential advantages of a segmentation strategy that is not exclusively reliant on specific cell-type characteristics.

The introduction of non-ambiguously labeled images into the model led to a sustained improvement in the model's performance, with no apparent diminishing returns. This indicates, that there is still room for improvement when adding more non-ambiguously labeled data.

Experiment 3 explored the potential of pseudo-labeling. Comparing a SSAL Model with an uncorrected pseudo-labeled model, it was evident that the SSAL Model not only enhanced model performance, but also decreased annotation time by over 60%. These findings emphasize the potential benefits of incorporating pseudo-labeling in the training process, reinforcing the vital role of manual correction in improving segmentation accuracy.

However, upon comparing the models from Experiments 2 and 3, it was apparent that while Experiment 2 outperformed Experiment 3 quantitatively, the SSAL Model from Experiment 3 surpassed the model from Experiment 2 in qualitative aspects. This variation could be due to differences in dataset composition, potential overfitting in the SSAL model, and the nature of the performance metrics used. This discrepancy emphasizes the importance of considering both qualitative and quantitative evaluations in model assessments.

The main insight derived from this research is the immense potential of semi-supervised active learning in significantly reducing manual annotation time. By integrating automated methods with human supervision, we create a synergy of the speed and efficiency of automated labeling with the accuracy and precision of manual correction. Should the pending Fpol evaluation by the experts indicate that the images segmented with our model achieve similar Fpol values, then the marked improvements in performance, alongside the 99.7% time saved in manual

segmentation, propose that this approach could revolutionize the diagnosis of thyroid cancer, thus serving the needs of thousands affected by this disease.

5.2 Future Work

Despite the promising results achieved thus far, it is evident that there are still opportunities for improvements in order to further enhance our model. To systematically address these areas of potential advancement, we have structured our proposed improvements and future research into three distinct timeframes: short term (within weeks), midterm (within months), and long term (longer than 6 months).

Short Term

Replacing Partially Labeled Data: In the short term, a significant improvement to the thyroid cell segmentation model could be achieved by expanding the dataset. The best-performing model was trained with a dataset of just 50 partially labeled, 16 fully labeled and 14 pseudo-labeled images, which, although the results were great, is relatively small for a deep learning model. Therefore, adding more non-ambiguously labeled images of thyroid cells could increase the model's generalization abilities, reducing overfitting and improving overall performance [28]. As our experiments have shown, adding non-ambiguously labeled data to the existing ambiguously labeled dataset can lead to a significantly better performance of the model. As there are no diminishing returns visible with increasing amount of fully labeled data added in Figure 4.5, there is potential for improvement when converting the 50 partially labeled images to fully labeled images.

Improving Post-Processing: To handle unusually segmented cells such as in Figure 4.7, we could refine our post-processing techniques. Drawing from the work by Kaushal and Singla [52], we could adopt their automated segmentation and post-processing methodology, which is context-sensitive, magnification independent, and requires no prior setting of parameters like the time step or weighted area coefficient. This approach also incorporates neighborhood information present in the image using an energy curve, which facilitates more favorable automatic threshold selection. This approach incorporates techniques such as area opening and dilation, which independently adjust their parameters based on the output of the segmentation. This adaptive approach allows for more accurate refinement of the segmentation results.

This automatic determination reduces the time spent on manual parameter tuning, and could be particularly beneficial in handling high-density cell clusters in our data. This way, our model can adaptively enhance segmentations, especially for tightly packed cell populations, which it currently struggles with. Incorporating this method could provide a more nuanced understanding of cellular structure and better performance overall.

Data Augmentation: To navigate the challenge of having a small dataset, data augmentation techniques could be a promising strategy. One such technique is the use of Perlin noise, a type of gradient noise developed by Ken Perlin. This method was notably employed in a recent study for augmenting high-resolution computed tomography (HRCT) images to classify patterns of diffuse interstitial lung disease (DILD) [30]. The results showed a significant increase in the accuracy of their deep learning model, a reduction in overfitting, and a wider variety of DILD disease patterns for the model's training.

The successful application of Perlin noise in the aforementioned study presents an intriguing possibility for cell segmentation models as well. Perlin noise could be used to generate diverse

synthetic training examples, creating randomness in cell shapes, textures, and distributions. This could potentially enhance our model's ability to generalize and perform better with unseen data. However, we have not found specific examples of Perlin noise being applied to cell image datasets as of yet. This highlights an opportunity for novel exploration and application of this technique within this context.

Midterm

Attention Mechanisms: In a potential future direction, we suggest investigating the inclusion of attention mechanisms, as demonstrated by the AURA-net model [53]. Attention mechanisms are a key feature of this model, enabling it to selectively focus on pertinent aspects of the image while filtering out less important details. The AURA-net model demonstrates impressive potential, even when relying on a small dataset. Utilizing a dataset of just about 30 training examples, the model managed to surpass the performance of state-of-the-art alternatives such as the U-Net. This strongly suggests that the integration of attention mechanisms could potentially lead to substantial improvements in our model's performance, even with limited training data.

Data Simulation Strategies: To alleviate the limitations of a small dataset in the midterm, The paper "Using simulated fluorescence cell micrographs for the evaluation of cell image segmentation algorithms" by Wiesmann et al. [54] presents a new approach for creating realistic simulations of fluorescent cell micrographs. This novel technique also inherently provides accurate ground truth, which circumvents the necessity for manual annotations that could potentially introduce errors.

Implementing this simulation method could lead to significant augmentation of our current dataset, thereby enabling our model to learn from a more diverse array of cell images. The additional simulated samples could effectively enhance both the diversity and size of our dataset, thereby reducing the risk of overfitting. The authors further highlight that their simulation technique can generate images with various complexity levels, including challenging scenarios with high texture variability and overlapping cells [54]. This could address some of our model's present limitations.

Another approach to generate more data is to use a two-stage generative adversarial network (GAN) as proposed by Pandey et al [55]. This method employs a first-stage GAN to generate a synthetic binary mask, and then a second-stage GAN uses this synthetic mask to produce the corresponding synthetic image. This strategy could serve as an effective means to artificially expand our dataset, producing additional image-mask pairs. It could ultimately enhance the performance of our image segmentation model, particularly in cases where we encounter a variety of segmentation complexities.

Incorporating this approach could provide an additional pathway for enhancing our model's performance. As noted by Pandey et al. [55], the benefit of their method starts to decrease when an extremely large number of synthetic images ($\geq$20K) are used for training. They suggest this decline could be attributed to the inclusion of low-quality synthetic images, such as those with heavily overlapping nuclei or inaccurate coloration [55]. This insight would need to be taken into account when implementing this strategy into our work to ensure optimal results.

Cost-Effective Active Learning: The incorporation of a CEAL strategy [46], a concept described in Chapter 2.6, could potentially enhance our SSAL approach, optimizing the process of expanding our labeled data while maintaining its quality.

CEAL synergizes active learning with an intuitive user interface that facilitates the user in correcting, adding, or deleting labels. This strategy enables our model to actively learn not only from newly encountered data but also from the iterative refinement and correction of pseudo-labeled data, which is a core element of our SSAL strategy. By promoting an interactive feedback loop between the user and the model, the CEAL approach could further improve the efficiency of the labeling process and accelerate the model's performance improvement [46].

The key advantage of integrating CEAL into our SSAL strategy lies in its ability to make the most of the strengths of both the model and the user. The model can identify areas where it requires the most guidance, and the user can provide targeted input to direct the model's learning. This approach could be particularly beneficial when our dataset is limited, enhancing the robustness and generalizability of our cell segmentation model, and further reducing the annotation time, a pivotal aspect of our SSAL strategy.

Exploring New Architectures: For future enhancements to our U-Net-based cell segmentation model, adopting methodologies from related studies may prove beneficial. A prime example is the work of Scherr et al. [56], titled "Cell segmentation and tracking using CNN-based distance predictions and a graph-based matching strategy". The proposed method in their paper leverages an innovative representation of cell borders via distance maps and an adapted U-Net convolutional neural network (CNN) with two decoder paths. This representation enables the model to learn not only from touching but also from closely situated cells, addressing a key limitation in our current model – the segmentation of clustered or overlapping cells. Scherr et al.'s [56] approach is highly robust to annotation errors and flexible enough to handle underrepresented or excluded cell types from the training data, which could improve our model's robustness and accuracy. Additionally, their results underscore the potential of specialized loss functions for improving segmentation results, particularly for the dominant cell types in the training set [56].

CycleGAN for Handling Partially Labeled Data: To address the challenge of ambiguously or partially labeled data in cell segmentation, we could utilize CycleGAN [57], a form of generative adversarial network that is particularly well-suited to image-to-image translation tasks. It learns to transform images from one domain to another, while preserving key features and characteristics. In our context, CycleGANs could be used to synthesize fully labeled cell images from those that are partially labeled [58]. This would be achieved by extracting labeled cells and their positional information from the partially labeled images and using this data to train the CycleGAN. According to the work of He et al. [58], by employing such an approach, it was possible to reduce the required manual labor for labeling by more than 50%, while achieving a similar recognition accuracy to that provided by experts. This approach could enrich our training dataset, helping us to mitigate the effects of incomplete labeling and potentially improving the performance of our segmentation model, as well as reduce manual annotation time.

Three-Stage Cancer Diagnosis Model: Improving cancer diagnosis speed and workflow could be achieved via a three-step process involving segmentation, feature extraction, and malignancy prediction. First, our segmentation model is employed for cell segmentation, distinguishing individual cells from surrounding tissue. Next, we extract relevant cell features, such as shape, size, texture, intensity, and Fpol. Finally, these extracted features are fed into a separate machine learning model trained to differentiate malignant from benign cells, thereby automating the prediction of malignancy. Incorporating this automated process into diagnostic workflows could significantly reduce diagnosis time and enhance accuracy.

Long Term

Data Diversity: In the long term, addressing the diversity of the dataset can significantly enhance the robustness and generalizability of our cell segmentation model [59]. One crucial aspect of this diversity is the variation in image capturing devices. Different devices can produce images with significant variations in brightness, contrast, color balance, and resolution [60]. These differences can pose substantial challenges for image segmentation models, potentially reducing their accuracy and generalizability. By expanding our dataset to include images from a range of different capturing devices, we can train our model to adapt to these variations. This practice is especially critical in medical imaging, where different hospitals and clinics may use different equipment. Therefore, investing in a long-term strategy to systematically collect samples from various image capturing devices could significantly improve the performance of our cell segmentation model, making it more robust and widely applicable across different imaging conditions.

Deep Adversarial Networks (DAN): To improve our cell segmentation model in the long term, incorporating DAN [61] could be a promising strategy. As demonstrated in the study by Zhang et al. [61], the unique capability of DAN to train using unannotated images can mitigate the common constraint of limited annotated images in biomedical image analysis, thereby broadening our training dataset.

DAN architecture consists of two core components: the segmentation network (SN) and the evaluation network (EN). These two components collaborate in an iterative adversarial training process, enhancing the segmentation accuracy over time. The EN scrutinizes the SN's segmentation outputs on unannotated images, offering critical feedback that guides the fine-tuning of the SN, resulting in continual improvement and adaptation of the model [61]. Through this adversarial process, the DAN learns to generate better segmentation maps, even when the original training images were unannotated. This is possible because the process essentially leverages the EN's ability to differentiate between "real" and "fake" (i.e. manually-annotated versus machine-generated) segmentation maps to guide the learning of the SN. Thus, unannotated images are indirectly used to train the SN [61]. Since manual annotation of cell images is resource-intensive, implementing DAN can help improve model efficiency and reduce manual annotation time by effectively utilizing unannotated images, thereby expanding our training dataset and potentially enhancing model robustness.

Fine-Tuning Segment Anything Model (SAM): SAM is a cutting-edge machine learning model for image segmentation tasks [62]. It employs the zero-shot learning paradigm, where the model is designed to handle classes it has not seen during training [63]. Despite the promising aspects of SAM, a study [64] indicates that it underperforms compared to state-of-the-art medical image segmentation algorithms. This underperformance is mostly observed in the context of medical image segmentation tasks.

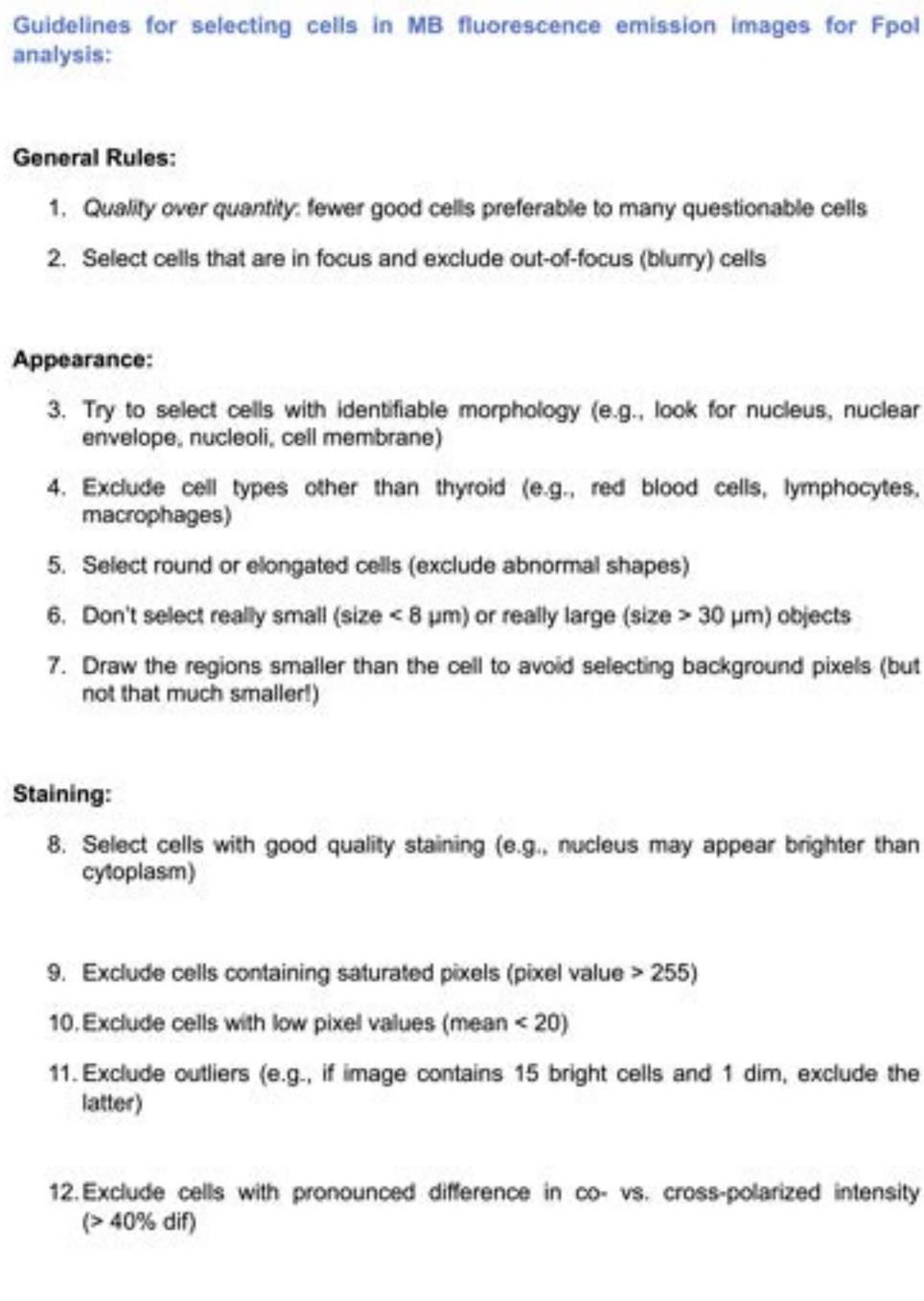
Nevertheless, recent research by Ma et al. [65] indicates that SAM can be fine-tuned for significant improvements across various segmentation tasks and image modalities. Despite these improvements, fine-tuned SAM still lags behind specialist models, particularly in boundary consensus, and struggles with objects that have missing boundaries, low contrast, or are tiny. It also tends to generate inaccurate segmentation results when many similar objects are inside the bounding box. However, these limitations can potentially be overcome by using larger models, increasing the dataset size, and leveraging other advancements such as scribble-based prompts for user interaction [65].

Furthermore, Ma et al. [65] suggest that integrating SAM into commonly used medical image

viewers will enable more end-users to access state-of-the-art models without extensive coding knowledge. They predict a "GPT-4 moment" for the field of medical image segmentation in the near future, where large-scale models similar to SAM become the foundation for the field. In conclusion, while there are certain areas for improvement, SAM shows promise in the medical image segmentation domain, providing a foundation for future research and development in the field [65].

A Appendix

A.1 Guidelines for selecting cells



Guidelines for selecting cells in MB fluorescence emission images for Fpol analysis:

General Rules:

1. *Quality over quantity*: fewer good cells preferable to many questionable cells
2. Select cells that are in focus and exclude out-of-focus (blurry) cells

Appearance:

3. Try to select cells with identifiable morphology (e.g., look for nucleus, nuclear envelope, nucleoli, cell membrane)
4. Exclude cell types other than thyroid (e.g., red blood cells, lymphocytes, macrophages)
5. Select round or elongated cells (exclude abnormal shapes)
6. Don't select really small (size < 8 µm) or really large (size > 30 µm) objects
7. Draw the regions smaller than the cell to avoid selecting background pixels (but not that much smaller!)

Staining:

8. Select cells with good quality staining (e.g., nucleus may appear brighter than cytoplasm)
9. Exclude cells containing saturated pixels (pixel value > 255)
10. Exclude cells with low pixel values (mean < 20)
11. Exclude outliers (e.g., if image contains 15 bright cells and 1 dim, exclude the latter)
12. Exclude cells with pronounced difference in co- vs. cross-polarized intensity (> 40% dif)

FIGURE A.1: Researcher Guidelines

A.2 Experiment results

Baseline Experiment (Experiment One)

TABLE A.1: Evaluation metrics for both the training and testing datasets in the baseline experiment. For the training dataset, the metrics include the loss value ("loss"), Intersection over Union score ("tr-iou"), Dice coefficient score ("tr-dic"), precision ("tr-pre"), recall ("tr-rec"), and F1-score ("tr-f1s"). Similarly, for the testing dataset, the metrics include the IoU score ("te-iou"), Dice coefficient score ("te-dic"), precision ("te-pre"), recall ("te-rec"), and F1-score ("te-f1s").

net	bs	lr	ep	loss	tr-iou	tr-dic	tr-pre	tr-rec	tr-f1s	te-iou	te-dic	te-pre	te-rec	te-f1s
modified	5	1e-03	500	0.130	0.522	0.671	0.726	0.667	0.695	0.394	0.528	***0.834***	0.444	0.580
modified	5	1e-03	1000	0.086	0.726	0.835	0.865	0.820	0.842	0.411	0.554	0.804	0.471	0.594
modified	5	1e-03	3000	0.051	0.812	0.892	0.915	0.878	0.896	0.413	0.559	0.778	0.479	0.593
modified	5	1e-05	500	0.291	0.654	0.783	0.806	0.782	0.794	0.425	0.569	0.741	0.521	0.612
modified	5	1e-05	1000	0.167	0.744	0.848	0.853	0.856	0.854	0.425	0.568	0.737	***0.524***	0.613
modified	5	1e-05	3000	0.043	0.823	0.897	0.920	0.887	***0.903***	0.428	0.577	0.747	0.515	0.610
modified	8	1e-03	500	0.140	0.529	0.676	0.743	0.657	0.697	0.442	0.593	0.827	0.496	***0.620***
modified	8	1e-03	1000	0.105	0.666	0.792	0.820	0.781	0.800	***0.448***	***0.598***	0.816	0.500	***0.620***
modified	8	1e-03	3000	***0.035***	***0.853***	***0.914***	***0.939***	***0.903***	0.921	0.361	0.503	0.821	0.407	0.544
modified	8	1e-05	500	0.298	0.615	0.753	0.794	0.741	0.767	0.421	0.572	0.756	0.496	0.599
modified	8	1e-05	1000	0.253	0.718	0.829	0.828	0.848	0.838	0.416	0.569	0.795	0.490	0.606
modified	8	1e-05	3000	0.062	0.815	0.892	0.914	0.884	0.898	0.402	0.555	0.780	0.464	0.582
full	5	1e-03	500	0.075	0.742	0.846	0.873	0.833	0.853	***0.414***	***0.558***	0.720	***0.508***	***0.596***
full	5	1e-03	1000	0.055	0.796	0.881	0.905	0.869	0.887	0.392	0.541	0.719	0.475	0.572
full	5	1e-03	3000	***0.037***	***0.851***	***0.914***	***0.938***	***0.902***	***0.920***	0.372	0.523	0.741	0.445	0.556
full	5	1e-05	500	0.277	0.165	0.276	0.262	0.473	0.337	0.180	0.298	0.290	0.411	0.340
full	5	1e-05	1000	0.234	0.171	0.280	0.427	0.267	0.328	0.202	0.322	0.466	0.282	0.351
full	5	1e-05	3000	0.193	0.157	0.242	0.763	0.183	0.295	0.228	0.341	0.793	0.259	0.390
full	8	1e-03	500	0.076	0.722	0.832	0.857	0.823	0.840	0.402	0.555	0.770	0.473	0.586
full	8	1e-03	1000	0.059	0.789	0.878	0.900	0.866	0.883	0.373	0.524	0.765	0.435	0.555
full	8	1e-03	3000	0.036	0.847	0.913	0.934	0.900	0.917	0.392	0.549	0.768	0.453	0.570
full	8	1e-05	500	0.301	0.159	0.268	0.239	0.517	0.327	0.162	0.271	0.314	0.381	0.344
full	8	1e-05	1000	0.256	0.171	0.285	0.293	0.417	0.344	0.163	0.274	0.354	0.308	0.330
full	8	1e-05	3000	0.213	0.093	0.152	0.781	0.103	0.182	0.085	0.139	***0.877***	0.089	0.161
partial	5	1e-03	500	0.153	0.456	0.611	0.705	0.589	0.642	0.401	0.544	0.709	0.491	0.581
partial	5	1e-03	1000	0.143	0.528	0.679	0.735	0.663	0.697	0.416	0.558	0.704	0.508	0.590
partial	5	1e-03	3000	0.102	***0.678***	***0.801***	***0.837***	***0.782***	***0.808***	***0.426***	***0.576***	0.754	***0.517***	***0.613***
partial	5	1e-05	500	0.333	0.164	0.275	0.262	0.502	0.344	0.182	0.303	0.277	0.466	0.348
partial	5	1e-05	1000	0.270	0.181	0.296	0.369	0.348	0.358	0.201	0.326	0.381	0.330	0.353
partial	5	1e-05	3000	0.233	0.175	0.268	0.732	0.209	0.325	0.237	0.354	0.775	0.270	0.401
partial	8	1e-03	500	0.156	0.443	0.598	0.700	0.573	0.630	0.375	0.524	0.719	0.455	0.558
partial	8	1e-03	1000	0.140	0.511	0.663	0.739	0.637	0.684	0.374	0.524	0.752	0.441	0.556
partial	8	1e-03	3000	***0.101***	0.666	0.793	0.827	0.775	0.800	0.404	0.559	0.788	0.472	0.590
partial	8	1e-05	500	0.361	0.156	0.264	0.243	0.567	0.340	0.170	0.284	0.319	0.464	0.378
partial	8	1e-05	1000	0.277	0.163	0.272	0.288	0.427	0.344	0.154	0.263	0.342	0.326	0.334
partial	8	1e-05	3000	0.246	0.155	0.243	0.712	0.185	0.293	0.154	0.241	***0.805***	0.171	0.282

Non-Ambiguous Data Experiment (Experiment Two)

TABLE A.2: Evaluation metrics for both the training and testing datasets in the non-ambiguous experiment. For the training dataset, the metrics include the loss value ("loss"), Intersection over Union score ("tr-iou"), Dice coefficient score ("tr-dic"), precision ("tr-pre"), recall ("tr-rec"), and F1-score ("tr-f1s"). Similarly, for the testing dataset, the metrics include the IoU score ("te-iou"), Dice coefficient score ("te-dic"), precision ("te-pre"), recall ("te-rec"), and F1-score ("te-f1s").

net	bs	lr	ep	loss	tr-iou	tr-dic	tr-pre	tr-rec	tr-f1s	te-iou	te-dic	te-pre	te-rec	te-f1s
modified	5	1e-03	500	0.105	0.737	0.845	0.845	0.856	0.850	0.402	0.557	0.556	0.657	0.602
modified	5	1e-03	1000	0.077	0.815	0.897	0.904	0.892	0.898	0.490	0.643	0.638	0.704	0.670
modified	5	1e-03	3000	0.036	0.869	0.930	0.936	0.924	0.930	0.501	0.650	0.657	0.726	0.689
modified	5	1e-05	500	0.221	0.778	0.873	0.860	0.893	0.876	0.502	0.658	0.630	0.742	0.681
modified	5	1e-05	1000	0.118	0.820	0.900	0.898	0.903	0.901	0.533	0.686	0.643	0.778	0.704
modified	5	1e-05	3000	0.041	0.863	0.926	0.931	0.921	0.926	***0.575***	0.720	***0.673***	0.814	0.737
modified	8	1e-03	500	0.093	0.770	0.868	0.868	0.873	0.871	0.443	0.605	0.554	0.755	0.639
modified	8	1e-03	1000	0.069	0.828	0.905	0.913	0.899	0.906	0.491	0.646	0.587	0.775	0.668
modified	8	1e-03	3000	***0.025***	***0.900***	***0.947***	***0.952***	***0.943***	***0.947***	0.574	***0.722***	0.637	***0.876***	***0.738***
modified	8	1e-05	500	0.227	0.788	0.880	0.862	0.903	0.882	0.513	0.673	0.574	0.846	0.684
modified	8	1e-05	1000	0.183	0.82	0.900	0.887	0.916	0.901	0.526	0.683	0.595	0.844	0.698
modified	8	1e-05	3000	0.041	0.868	0.929	0.935	0.924	0.929	0.559	0.711	0.619	0.871	0.724
full	5	1e-03	500	0.073	0.815	0.897	0.902	0.893	0.898	0.490	0.640	0.605	0.715	0.656
full	5	1e-03	1000	0.057	0.837	0.911	0.917	0.906	0.912	0.517	0.664	0.621	0.748	0.679
full	5	1e-03	3000	0.032	0.874	0.933	0.938	0.928	0.933	***0.562***	0.706	***0.643***	0.831	***0.725***
full	5	1e-05	500	0.224	0.321	0.476	0.360	0.778	0.492	0.205	0.335	0.221	0.788	0.346
full	5	1e-05	1000	0.228	0.434	0.587	0.700	0.556	0.619	0.303	0.450	0.452	0.516	0.482
full	5	1e-05	3000	0.160	0.484	0.620	0.857	0.552	0.671	0.313	0.443	***0.643***	0.475	0.546
full	8	1e-03	500	0.058	0.820	0.901	0.907	0.895	0.901	0.523	0.680	0.594	0.827	0.691
full	8	1e-03	1000	0.056	0.846	0.916	0.923	0.909	0.916	0.536	0.688	0.593	0.850	0.699
full	8	1e-03	3000	***0.031***	***0.880***	***0.936***	***0.940***	***0.932***	***0.936***	0.560	***0.708***	0.606	***0.882***	0.718
full	8	1e-05	500	0.266	0.305	0.457	0.335	0.782	0.470	0.188	0.312	0.203	0.776	0.321
full	8	1e-05	1000	0.234	0.372	0.536	0.461	0.688	0.552	0.235	0.378	0.269	0.688	0.387
full	8	1e-05	3000	0.165	0.326	0.451	0.898	0.346	0.500	0.227	0.346	0.629	0.302	0.408
partial	5	1e-03	500	0.138	0.710	0.826	0.821	0.842	0.832	0.426	0.584	0.543	0.686	0.606
partial	5	1e-03	1000	0.096	0.753	0.856	0.856	0.862	0.859	0.461	0.619	0.583	0.717	0.643
partial	5	1e-03	3000	0.060	0.812	0.895	0.904	0.887	0.896	0.511	0.667	0.612	0.780	0.686
partial	5	1e-05	500	0.268	0.365	0.525	0.402	0.829	0.542	0.221	0.356	0.235	***0.825***	0.365
partial	5	1e-05	1000	0.217	0.462	0.623	0.560	0.757	0.644	0.295	0.450	0.338	0.728	0.462
partial	5	1e-05	3000	0.222	0.488	0.619	0.838	0.564	0.674	0.311	0.438	***0.661***	0.467	0.547
partial	8	1e-03	500	0.116	0.706	0.824	0.822	0.838	0.830	0.419	0.582	0.510	0.725	0.599
partial	8	1e-03	1000	0.095	0.765	0.865	0.868	0.867	0.867	0.460	0.621	0.550	0.755	0.636
partial	8	1e-03	3000	***0.067***	***0.819***	***0.900***	***0.907***	***0.894***	***0.900***	***0.518***	***0.676***	0.595	0.823	***0.690***
partial	8	1e-05	500	0.322	0.338	0.493	0.374	0.801	0.510	0.206	0.338	0.220	0.801	0.345
partial	8	1e-05	1000	0.267	0.396	0.557	0.433	0.849	0.574	0.230	0.370	0.245	0.822	0.377
partial	8	1e-05	3000	0.207	0.431	0.559	0.844	0.492	0.621	0.288	0.425	0.559	0.440	0.492

A.3 API

The project, oswald-martin/zhaw_ba_cell_segmentation is organized into four main directories: ' app', 'backend', 'frontend', and 'malearn'. Each directory plays a vital role in the functioning and structure of the application.

- **app:** serves as the application's core. It contains the 'main.py' script, which is the entry point for the application. This script is responsible for orchestrating the server and worker processes. These processes follow a producer-consumer pattern central to the application's working. The script creates these processes, starts them, and ensures they terminate correctly.

- **backend:** Forms the server-side part of the API. It holds the server and worker processes and utility functions that perform image segmentation and serve the React frontend. The server process, also referred to as the producer, segments the images and temporarily stores the output. The worker process, or the consumer, removes the output from the temporary storage after a specified duration. The 'backend' directory further consists of subdirectories that hold scripts for API routes, a communication queue between the server and worker, the server process, utility functions for API calls, and the worker process.

- **frontend:** Provides a user-friendly platform for processing images. Built using Vite, TypeScript, React, and Tailwind CSS, this interface is designed for intuitive interaction. It communicates seamlessly with the FastAPI backend and provides an upload and download interface for users to input and receive segmented cell images. The directory houses source files for the React application under the 'src' directory, the output directory 'dist' for the 'npm run build' command, and the 'index.html' file which serves as the container for the React app.

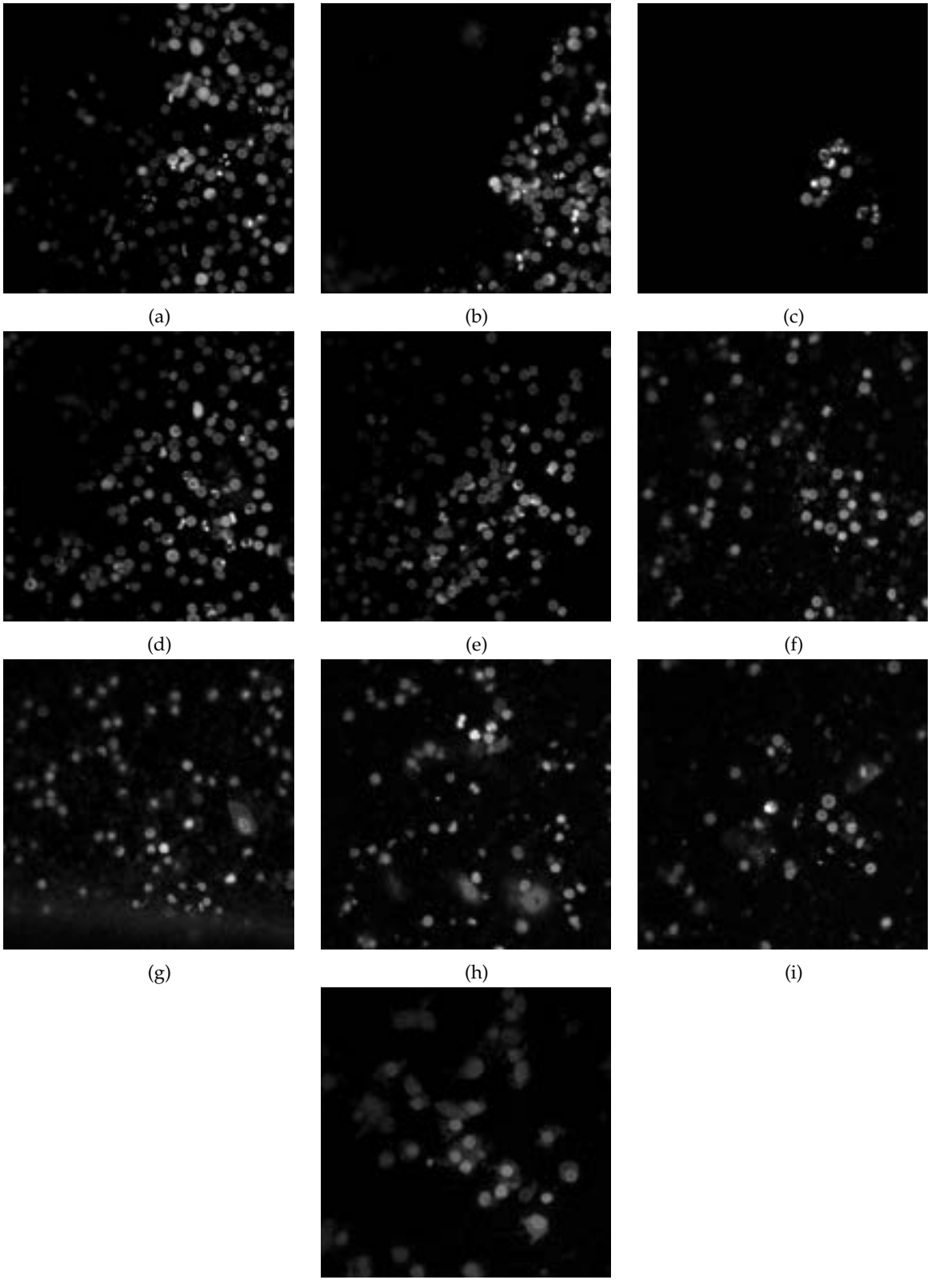
- **malearn:** Encapsulates the machine learning part of the project. It includes dataset handling, implementation of the U-Net model, and various utility functions. The directory is divided into several subdirectories, including 'config' which contains configuration files for different models, 'data' for dataset loading and generation of training and validation data, 'models' for the implementation of the U-Net model, and 'util' for providing utility functions for training and evaluation.

The project can also be built and run using Docker. The Docker image can be pulled directly from DockerHub using the command 'docker pull oswaldmartin/cellseg'. To run the Docker container with the required environment variables, the 'docker run' command is used with the '-e' flag to set these variables.

The environment variables include:

- `MODEL_NAME`: Selects the model ("normal", "trans", "freeze").
- `MODEL_CONF`: Sets the pixel confidence (float between 0.0 and 1).
- `DRAW_CONTOURS`: Toggles segmentation contours (0 or 1).
- `INFO_FILE`: Provides execution info (0 or 1).

A.4 Qualitative Evaluation Images



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j)

FIGURE A.2: (a) sample 1 (b) sample 2 (c) sample 3 (d) sample 4 (e) sample 5 (f) sample 6 (g) sample 7 (h) sample 8 (i) sample 9 (j) sample 10

Bibliography

[1] G. Grani, M. Sponziello, V. Pecce, V. Ramundo, and C. Durante, "Contemporary thyroid nodule evaluation and management," *Journal of Clinical Endocrinology and Metabolism*, vol. 105, no. 9, pp. 2869–2883, Sep. 2020, ISSN: 19457197. DOI: 10.1210/CLINEM/DGAA322.

[2] *Thyroid Statistics | American Cancer Society - Cancer Facts & Statistics*. [Online]. Available: https://cancerstatisticscenter.cancer.org/?_ga=2.206352275.102032163.1681126140-2041538640.1681126140#!/cancer-site/Thyroid.

[3] P. R. Jermain, A. H. Fischer, L. Joseph, A. Muzikansky, and A. N. Yaroslavsky, "Fluorescence Polarization Imaging of Methylene Blue Facilitates Quantitative Detection of Thyroid Cancer in Single Cells," *Cancers*, vol. 14, no. 5, Mar. 2022, ISSN: 2072-6694. DOI: 10.3390/CANCERS14051339. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/35267647/.

[4] F. Fulciniti, A. Cipolletta Campanile, M. G. Malzone, *et al.*, "Impact of ultrasonographic features, cytomorphology and mutational testing on malignant and indeterminate thyroid nodules on diagnostic accuracy of fine needle cytology samples: A prospective analysis of 141 patients," *Clinical Endocrinology*, vol. 91, no. 6, p. 851, Dec. 2019, ISSN: 13652265. DOI: 10.1111/CEN.14089. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6972562/.

[5] C. Mathiot, D. Decaudin, J. Klijanienko, *et al.*, "Fine-needle aspiration cytology combined with flow cytometry immunophenotyping is a rapid and accurate approach for the evaluation of suspicious superficial lymphoid lesions," *Diagnostic Cytopathology*, vol. 34, no. 7, pp. 472–478, Jul. 2006, ISSN: 87551039. DOI: 10.1002/dc.20487.

[6] R. Paschke, S. Cantara, A. Crescenzi, B. Jarzab, T. J. Musholt, and M. S. Simoes, "European Thyroid Association Guidelines regarding Thyroid Nodule Molecular Fine-Needle Aspiration Cytology Diagnostics," *European Thyroid Journal*, vol. 6, no. 3, p. 115, 2017, ISSN: 2235-0640. DOI: 10.1159/000468519. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5527175/.

[7] M. Vaiman, A. Nagibin, and J. Olevson, "Complications in primary and completed thyroidectomy," *Surgery Today*, vol. 40, no. 2, pp. 114–118, Feb. 2010, ISSN: 09411291. DOI: 10.1007/S00595-008-4027-9. [Online]. Available: https://link.springer.com/article/10.1007/s00595-008-4027-9.

[8] A. N. Yaroslavsky, X. Feng, A. Muzikansky, and M. R. Hamblin, "Fluorescence Polarization of Methylene Blue as a Quantitative Marker of Breast Cancer at the Cellular Level," *Scientific Reports*, vol. 9, no. 1, p. 940, Jan. 2019, ISSN: 2045-2322. DOI: 10.1038/s41598-018-38265-0.

[9] E. Meijering, "Cell Segmentation: 50 Years Down the Road [Life Sciences]," *IEEE Signal Processing Magazine*, vol. 29, no. 5, pp. 140–145, Sep. 2012, ISSN: 1053-5888. DOI: 10.1109/MSP.2012.2204190.

[10] J. C. Caicedo, J. Roth, A. Goodman, *et al.*, "Evaluation of Deep Learning Strategies for Nucleus Segmentation in Fluorescence Images," *Cytometry Part A*, vol. 95, no. 9, pp. 952–965, Sep. 2019, ISSN: 1552-4922. DOI: 10.1002/cyto.a.23863.

[11] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[12] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, Jan. 2015, ISSN: 08936080. DOI: 10.1016/j.neunet.2014.09.003.

[13] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, ISSN: 0028-0836. DOI: 10.1038/nature14539.

[14] Alexander LeNail, *NN SVG*. [Online]. Available: http://alexlenail.me/NN-SVG/index.html.

[15] Educative, *Overfitting and underfitting*. [Online]. Available: https://www.educative.io/answers/overfitting-and-underfitting.

[16] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2274–2281, 2012, ISSN: 01628828. DOI: 10.1109/TPAMI.2012.120.

[17] W. Khan, "Image Segmentation Techniques: A Survey," *Journal of Image and Graphics*, vol. 1, pp. 166–170, May 2014. DOI: 10.12720/joig.1.4.166-170.

[18] J. Long, E. Shelhamer, and T. Darrell, *Fully Convolutional Networks for Semantic Segmentation*, 2015. DOI: 10.48550/arXiv.1411.4038.

[19] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," May 2015.

[20] K. He, G. Gkioxari, P. Dollár, and R. Girshick, *Mask R-CNN*, 2018.

[21] J. Wu, W. Liu, C. Li, *et al.*, *A State-of-the-art Survey of U-Net in Microscopic Image Analysis: from Simple Usage to Structure Mortification*, 2022.

[22] Z. Zhang, C. Wu, S. Coleman, and D. Kerr, "DENSE-INception U-net for medical image segmentation," *Computer Methods and Programs in Biomedicine*, vol. 192, p. 105 395, Aug. 2020, ISSN: 0169-2607. DOI: 10.1016/J.CMPB.2020.105395.

[23] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, "U-Net and Its Variants for Medical Image Segmentation: A Review of Theory and Applications," *IEEE Access*, vol. 9, pp. 82 031–82 057, 2021, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2021.3086020.

[24] N. Abraham and N. M. Khan, "A Novel Focal Tversky loss function with improved Attention U-Net for lesion segmentation," Oct. 2018.

[25] K. Kawagoe, K. Hatano, S. Murakami, H. Lu, H. Kim, and T. Aoki, "Automatic Segmentation Method of Phalange Regions Based on Residual U-Net and MSGVF Snakes," in *2019 19th International Conference on Control, Automation and Systems (ICCAS)*, IEEE, Oct. 2019, pp. 1046–1049, ISBN: 978-89-93215-17-5. DOI: 10.23919/ICCAS47443.2019.8971740.

[26] R. Zhao, W. Chen, and G. Cao, "Edge-Boosted U-Net for 2D Medical Image Segmentation," *IEEE Access*, vol. 7, pp. 171 214–171 222, 2019, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2019.2953727.

[27] B. Wang, S. Jin, Q. Yan, *et al.*, "AI-assisted CT imaging analysis for COVID-19 screening: Building and deploying a medical AI system," *Applied Soft Computing*, vol. 98, p. 106 897, Jan. 2021, ISSN: 15684946. DOI: 10.1016/j.asoc.2020.106897.

[28] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, *Revisiting Unreasonable Effectiveness of Data in Deep Learning Era*, 2017.

[29] L. Perez and J. Wang, "The Effectiveness of Data Augmentation in Image Classification using Deep Learning," Dec. 2017. [Online]. Available: https://arxiv.org/abs/1712.04621v1.

[30] H. J. Bae, C. W. Kim, N. Kim, *et al.*, "A Perlin Noise-Based Augmentation Strategy for Deep Learning with Small Data Samples of HRCT Images," *Scientific Reports 2018 8:1*, vol. 8, no. 1, pp. 1–7, Dec. 2018, ISSN: 2045-2322. DOI: 10.1038/s41598-018-36047-2. [Online]. Available: https://www.nature.com/articles/s41598-018-36047-2.

[31] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges," *Journal of Digital Imaging*, vol. 32, no. 4, pp. 582–596, Aug. 2019, ISSN: 1618727X. DOI: 10.1007/S10278-019-00227-

X/TABLES/2. [Online]. Available: https://link.springer.com/article/10.1007/s10278-019-00227-x.

[32] P. Chlap, H. Min, N. Vandenberg, J. Dowling, L. Holloway, and A. Haworth, "A review of medical image data augmentation techniques for deep learning applications," *Journal of Medical Imaging and Radiation Oncology*, vol. 65, no. 5, pp. 545–563, Aug. 2021, ISSN: 1754-9477. DOI: 10.1111/1754-9485.13261.

[33] C.-K. Shie, C.-H. Chuang, C.-N. Chou, M.-H. Wu, and E. Y. Chang, "Transfer representation learning for medical image analysis," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, Aug. 2015, pp. 711–714, ISBN: 978-1-4244-9271-8. DOI: 10.1109/EMBC.2015.7318461.

[34] S. Khan, N. Islam, Z. Jan, I. Ud Din, and J. J. P. C. Rodrigues, "A novel deep learning based framework for the detection and classification of breast cancer using transfer learning," *Pattern Recognition Letters*, vol. 125, pp. 1–6, 2019, ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2019.03.022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167865519301059.

[35] H.-C. Shin, H. R. Roth, M. Gao, *et al.*, "Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1285–1298, May 2016, ISSN: 0278-0062. DOI: 10.1109/TMI.2016.2528162.

[36] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" *CoRR*, vol. abs/1411.1792, 2014. [Online]. Available: http://arxiv.org/abs/1411.1792.

[37] W. Lingyun, X. Yang, S. Li, T. Wang, P.-A. Heng, and D. Ni, *Cascaded Fully Convolutional Networks for automatic prenatal ultrasound image segmentation*. Apr. 2017, pp. 663–666. DOI: 10.1109/ISBI.2017.7950607.

[38] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, *et al.*, "Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?" *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1299–1312, May 2016, ISSN: 0278-0062. DOI: 10.1109/TMI.2016.2535302.

[39] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Deep End2End Voxel2Voxel Prediction," *CoRR*, vol. abs/1511.06681, 2015. [Online]. Available: http://arxiv.org/abs/1511.06681.

[40] N. Simmler, P. Sager, P. Andermatt, *et al.*, "A Survey of Un-, Weakly-, and Semi-Supervised Learning Methods for Noisy, Missing and Partial Labels in Industrial Vision Applications," *Proceedings - 2021 8th Swiss Conference on Data Science, SDS 2021*, pp. 26–31, Jun. 2021. DOI: 10.1109/SDS51136.2021.00012.

[41] L. Ye, Z. Liu, and Y. Wang, "Learning Semantic Segmentation with Diverse Supervision," *Proceedings - 2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018*, vol. 2018-January, pp. 1461–1469, Feb. 2018. DOI: 10.1109/WACV.2018.00164. [Online]. Available: https://arxiv.org/abs/1802.00509v1.

[42] I. Triguero, S. García, and F. Herrera, "Self-labeled techniques for semi-supervised learning: Taxonomy, software and empirical study," *Knowledge and Information Systems*, vol. 42, no. 2, pp. 245–284, Feb. 2015, ISSN: 02193116. DOI: 10.1007/S10115-013-0706-Y/FIGURES/13. [Online]. Available: https://link.springer.com/article/10.1007/s10115-013-0706-y.

[43] L. Yang, W. Zhuo, L. Qi, Y. Shi, and Y. Gao, "ST++: Make Self-training Work Better for Semi-supervised Semantic Segmentation," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2022-June, pp. 4258–4267, Jun. 2021, ISSN: 10636919. DOI: 10.1109/CVPR52688.2022.00423. [Online]. Available: https://arxiv.org/abs/2106.05095v2.

[44] O. T. Nartey, G. Yang, J. Wu, and S. K. Asare, "Semi-Supervised Learning for Fine-Grained Classification with Self-Training," *IEEE Access*, vol. 8, pp. 2109–2121, 2020, ISSN: 21693536. DOI: 10.1109/ACCESS.2019.2962258.

[45] M. Gorriz, A. Carlier, E. Faure, and X. Giro-i-Nieto, "Cost-Effective Active Learning for Melanoma Segmentation," Nov. 2017. [Online]. Available: https://arxiv.org/abs/1711.09168v2.

[46] K. Wang, D. Zhang, Y. Li, R. Zhang, and L. Lin, "Cost-Effective Active Learning for Deep Image Classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 12, pp. 2591–2600, Jan. 2017, ISSN: 10518215. DOI: 10.1109/tcsvt.2016.2589879. [Online]. Available: https://arxiv.org/abs/1701.03551v1.

[47] S.-j. Huang, R. Jin, and Z.-H. Zhou, "Active Learning by Querying Informative and Representative Examples," in *Advances in Neural Information Processing Systems*, J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, Eds., vol. 23, Curran Associates, Inc., 2010. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2010/file/5487315b1286f907165907aa8fc96619-Paper.pdf.

[48] J. Roels and Y. Saeys, "Cost-efficient segmentation of electron microscopy images using active learning," *CEUR Workshop Proceedings*, vol. 2491, Nov. 2019, ISSN: 16130073. [Online]. Available: https://arxiv.org/abs/1911.05548v1.

[49] J. C. Caicedo, A. Goodman, K. W. Karhohs, *et al.*, "Nucleus segmentation across imaging experiments: the 2018 Data Science Bowl," *Nature Methods*, vol. 16, no. 12, pp. 1247–1253, Dec. 2019, ISSN: 1548-7091. DOI: 10.1038/s41592-019-0612-7.

[50] Torchvision, *TorchVision: PyTorch's Computer Vision library*, 2016. [Online]. Available: https://github.com/pytorch/vision.

[51] S. Jiao, Y. Gao, J. Feng, T. Lei, and X. Yuan, "Does deep learning always outperform simple linear regression in optical imaging?" *Optics express*, vol. 28 3, pp. 3717–3731, 2019.

[52] C. Kaushal and A. Singla, "Automated segmentation technique with self-driven post-processing for histopathological breast cancer images," *CAAI Transactions on Intelligence Technology*, vol. 5, no. 4, pp. 294–300, Dec. 2020, ISSN: 2468-2322. DOI: 10.1049/TRIT.2019.0077. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1049/trit.2019.0077%20https://onlinelibrary.wiley.com/doi/abs/10.1049/trit.2019.0077%20https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/trit.2019.0077.

[53] E. Cohen and V. Uhlmann, "Aura-Net: Robust Segmentation Of Phase-Contrast Microscopy Images with Few Annotations," *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 640–644, 2021.

[54] V. Wiesmann, M. Bergler, R. Palmisano, M. Prinzen, D. Franz, and T. Wittenberg, "Using simulated fluorescence cell micrographs for the evaluation of cell image segmentation algorithms," *BMC Bioinformatics*, vol. 18, no. 1, pp. 1–12, Mar. 2017, ISSN: 14712105. DOI: 10.1186/S12859-017-1591-2/FIGURES/8. [Online]. Available: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-017-1591-2.

[55] S. Pandey, P. R. Singh, and J. Tian, "An image augmentation approach using two-stage generative adversarial network for nuclei image segmentation," *Biomed. Signal Process. Control.*, vol. 57, 2020.

[56] T. Scherr, K. Loeffler, M. Boehland, and R. Mikut, "Cell Segmentation and Tracking using CNN-Based Distance Predictions and a Graph-Based Matching Strategy," *CoRR*, vol. 2004.01486, 2020. [Online]. Available: https://arxiv.org/abs/2004.01486.

[57] J. .-. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2242–2251, ISBN: 2380-7504. DOI: 10.1109/ICCV.2017.244.

[58] J. He, C. Wang, D. Jiang, Z. Li, Y. Liu, and T. Zhang, "CycleGAN With an Improved Loss Function for Cell Detection Using Partly Labeled Images," *IEEE Journal of Biomedical and*

Health Informatics, vol. 24, no. 9, pp. 2473–2480, 2020, ISSN: 2168-2208. DOI: 10.1109/JBHI.2020.2970091.

[59] J. Krishnan, H. Purohit, and H. Rangwala, "Diversity-Based Generalization for Neural Unsupervised Text Classification under Domain Shift," *ArXiv*, vol. abs/2002.10937, 2020.

[60] G. M. De Luca, R. M. Breedijk, R. A. Brandt, *et al.*, "Re-scan confocal microscopy: scanning twice for better resolution.," *Biomedical optics express*, vol. 4, no. 11, p. 2644, Nov. 2013, ISSN: 2156-7085. DOI: 10.1364/BOE.4.002644.

[61] Y. Zhang, L. Yang, J. Chen, M. Fredericksen, D. P. Hughes, and D. Z. Chen, "Deep adversarial networks for biomedical image segmentation utilizing unannotated images," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10435 LNCS, pp. 408–416, 2017, ISSN: 16113349. DOI: 10.1007/978-3-319-66179-7{_}47/TABLES/2. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-66179-7_47.

[62] A. Kirillov, E. Mintun, N. Ravi, *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.

[63] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 9, pp. 2251–2265, 2018.

[64] S. He, R. Bao, J. Li, P. E. Grant, and Y. Ou, "Accuracy of segment-anything model (sam) in medical image segmentation tasks," *arXiv preprint arXiv:2304.09324*, 2023.

[65] J. Ma and B. Wang, "Segment anything in medical images," *arXiv preprint arXiv:2304.12306*, 2023.