Leveraging Weak Supervision for Cell Localization in Digital Pathology Using Multitask Learning and Consistency Loss

by

Berke Levent Cesur

A Dissertation Submitted to the
Graduate School of Sciences and Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Master of Science

in

Computer Sciences and Engineering

January 7, 2025

Leveraging Weak Supervision for Cell Localization in Digital Pathology Using Multitask Learning and Consistency Loss

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a master's thesis by

Berke Levent Cesur

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

_____

Prof. Çiğdem Gündüz-Demir (Advisor)

_____

Prof. Yücel Yemez

_____

Prof. Behçet Uğur Töreyin

Date: _____

to my mother and father...

ABSTRACT

Leveraging Weak Supervision for Cell Localization in Digital Pathology Using Multitask Learning and Consistency Loss
Berke Levent Cesur
Master of Science in Computer Sciences and Engineering
January 7, 2025

Cell detection and segmentation are integral parts of automated systems in digital pathology. Encoder-decoder networks have emerged as a promising solution for these tasks. However, training of these networks has typically required full boundary annotations of cells, which are labor-intensive and difficult to obtain on a large scale. However, in many applications, such as cell counting, weaker forms of annotations–such as point annotations or approximate cell counts–can provide sufficient supervision for training. This thesis proposes a new mixed-supervision approach for training multitask networks in digital pathology by incorporating cell counts derived from the eyeballing process–a quick visual estimation method commonly used by pathologists. This thesis has two main contributions: (1) It proposes a mixed-supervision strategy for digital pathology that utilizes cell counts obtained by eyeballing as an auxiliary supervisory signal to train a multitask network for the first time. (2) This multitask network is designed to concurrently learn the tasks of cell counting and cell localization, and this thesis introduces a consistency loss that regularizes training by penalizing inconsistencies between the predictions of these two tasks. Our experiments on two datasets of hematoxylin-eosin stained tissue images demonstrate that the proposed approach effectively utilizes the weakest form of annotation, improving performance when stronger annotations are limited. These results highlight the potential of integrating eyeballing-derived ground truths into the network training, reducing the need for resource-intensive annotations.

ÖZETÇE

Yüksek Lisans Tez Başlığı
Berke Levent Cesur
Bilgisayar Bilimleri ve Mühendisliği, Yüksek Lisans
7 Ocak 2025

Hücre tespiti ve segmentasyonu, dijital patolojide otomatik sistemlerin önemli bileşenleridir. Kodlayıcı-çözücü ağ modelleri, bu görevler için umut vaat eden bir çözüm olarak ortaya çıkmıştır. Ancak, bu tür ağların eğitimi genellikle hücrelerin kapsamlı sınır anotasyonlarını gerektirmektedir. Bu anotasyonları oluşturmak oldukça zahmetli olup büyük ölçekli veri olarak elde edilmesi zordur. Bununla birlikte, hücre sayımı gibi birçok uygulamada, zayıf anotasyon türleri – örneğin nokta anotasyonları veya yaklaşık hücre sayıları – eğitimi sağlamak için yeterli gözetimi sunabilir. Bu tez, hücre sayılarının patologların sıklıkla kullandığı hızlı bir gözle tarama ve yakınsama tekniği olan göz kararı yönteminden türetilmesiyle elde edilen zayıf anotasyonları içeren yeni bir karma gözetimli öğrenme yaklaşımı önermektedir. Bu tezin iki ana katkısı vardır: (1) İlk defa, göz kararı yöntemiyle elde edilen hücre sayılarının yardımcı bir gözetim sinyali olarak kullanıldığı karma gözetimli bir strateji önermektedir. (2) Önerilen çok görevli ağ, hücre sayımı ve hücre lokalizasyonu görevlerini eşzamanlı olarak öğrenmek üzere tasarlanmıştır ve bu tez, bu iki görevin tahminleri arasındaki tutarsızlıkları cezalandırarak eğitimi düzenleyen bir tutarlılık kayıp terimi önermektedir. Hematoksilin-eozin boyalı doku görüntülerinden oluşan iki veri kümesi üzerinde yapılan deneyler, önerilen yöntemin bu en zayıf anotasyon türünü etkili bir şekilde kullanarak, güçlü anotasyonların sınırlı olduğu durumlarda performansı iyileştirdiğini göstermektedir. Bu sonuçlar, ağ eğitimi sürecine göz kararı yöntemi ile türetilmiş gerçek değerlerin dahil edilme potansiyelini ortaya koymakta ve elde edilmesi yoğun kaynak kullanımı gerektiren anotasyonlara olan ihtiyacı azaltmaktadır.

ACKNOWLEDGMENTS

First and foremost, I would like to express my sincere gratitude to my advisor, Prof. Çiğdem Gündüz-Demir, for her unwavering guidance, support and patience throughout this research journey. Her deep expertise, insightful advice, and encouragement have been invaluable in helping me navigate every step of the thesis process.

I would also like to extend my heartfelt thanks to my thesis committee members, Prof. Yücel Yemez and Prof. Behçet Uğur Töreyin, for taking the time to review my work and provide invaluable feedback that significantly enhanced the quality of this study.

I am also grateful to Dr. Can Fahrettin Koyuncu for his continuous support and guidance and Assoc. Prof. İbrahim Kulaç, Dr. Ayşe Hümeyra Dur Karasayar, Dr. Çisel Aydın Meriçöz, Assoc. Prof. Pınar Bulutay, Prof. Nilgün Kapucuoğlu, Dr. Handan Eren, Dr. Ömer Faruk Dilbaz, Javidan Osmanli, Burhan Soner Yetkili for generously sharing the medical data and their expert knowledge, which were crucial to the success of this research.

I am deeply grateful to Koç University for providing an exceptional academic and research environment. My sincere thanks go to the esteemed faculty members and administrative staff, whose support and dedication have contributed immensely to my learning experience. I also acknowledge the financial support provided by the Scientific and Technological Research Council of Turkey (TÜBİTAK) under the project number 121E080 and the KUIS AI Center which made this study possible.

Lastly, I wish to express my profound gratitude to my family and friends for their constant love, understanding, and encouragement. Their unwavering belief in me has been a source of strength and motivation, enabling me to overcome challenges and reach this important milestone.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

In digital pathology, the development of automated decision support systems is essential to enable high-throughput analysis and reduce intra- and inter-observer variability. Typically, automated cell segmentation/detection is an integral part of these systems. They use the outputs of the segmentation/detection systems either directly, e.g., to count cells in a tissue, or as inputs to their subsequent steps, e.g., to construct a graph over the cells and run a graph neural network classifier. The most recent approach to localizing cells in a tissue image is to develop dense prediction networks [Ronneberger et al., 2015, Chen et al., 2021a, Fujita and Han, 2020]. When these networks are developed for segmentation, they require strong boundary annotations of the cells in the training data. The amount of the required data increases even more for training large-scale segmentation models [Graham et al., 2019, Greenwald et al., 2022]. On the other hand, delineating cell boundaries in a tissue image is one of the most labor-intensive tasks in digital pathology and obtaining such boundary annotations for a large number of cells is quite challenging. Nevertheless, although these segmentation networks, and thus this type of annotation, are necessary when one needs to extract histomorphological features from each cell separately, for many applications such as cell counting, it is sufficient to detect rough locations of the cells. Thus, weaker forms of annotation may be sufficient to train networks for such applications. Moreover, these weak annotations can be used in conjunction with strong annotations to better train the segmentation networks.

In the field of cell microscopy imaging, researchers have explored the use of a variety of weak annotations to train segmentation/detection networks in a weakly- or mixed-supervised learning scheme. These include scribbles [Oh et al., 2022], bound-

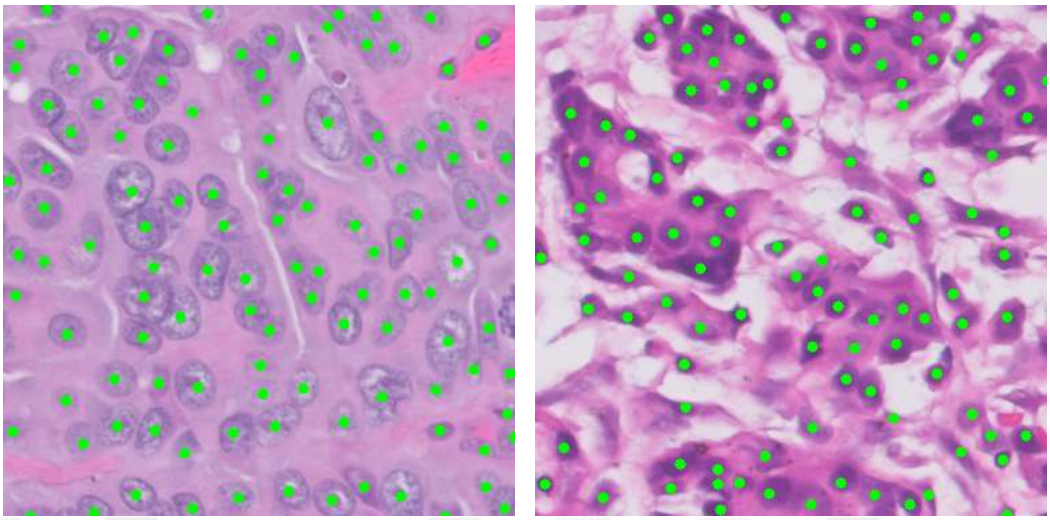


Figure 1.1: Examples tissue images from our in-house serous carcinoma dataset. Green dots were added to represent point annotations provided by our pathologist collaborators.

ing boxes [Aldughayfiq et al., 2023], but most commonly point annotations [Khalid et al., 2023, Miao et al., 2021], which involve placing a single point inside each cell (Figure 1.1). There are several studies in the literature that developed alternative training techniques to utilize point annotations. The studies in [Yu, 2023] and [Zhao and Yin, 2021] proposed to use self- and co-training techniques to generate cell segmentation masks from single point annotations. The latter study [Zhao and Yin, 2021] also allowed to keep human in the loop to provide additional annotations for finetuning the trained network. Other studies [Pan et al., 2018, Xie et al., 2018] localized cells in an image by training regression networks to learn density maps of the single point annotations. Different than the use of strong cell annotations to train a dense prediction network, these regression models necessitated preprocessing the point annotations using various techniques such as Gaussian kernels [Qu et al., 2020] and repel coding [Chamanzar and Nie, 2020]. The regression-based models were particularly advantageous in cases involving overlapping cells in densely populated cultures with background clutter.

In addition to cell localization, another important practical application of digital

pathology is to estimate the number of cells in a tissue region, which can serve as a significant indicator, e.g., clinicians may infer a patient's cancer stage or type by monitoring variations in cell counts [Chow et al., 2020, Frei et al., 2023]. The output of segmentation networks and regression-based models can also be used to estimate the cell counts. For example, the number of connected components in a map estimated by a segmentation network or the sum of pixel values in a density map estimated by a regression network can be employed to infer the cell counts. Alternatively, the cell count in an image can directly be used as a supervisory signal to train the models. In the literature, the cell count was used either as the sole ground truth to train a convolutional neural network [Xue et al., 2016, Lavitt et al., 2021] or as an auxiliary supervisor alongside the cell segmentation task in a multitask network [Hagos et al., 2019, Huang et al., 2022]. The multitask network defines a prediction branch for each task separately and learns the multiple tasks jointly from shared representations. Because of this joint learning to optimize shared parameters, multitask learning is not only known to be effective in improving the generalization ability of the model [Graham et al., 2023, Alom et al., 2020] but also has great potential in data utilization, especially when annotated data is scarce for one task but abundant for another, attribute that makes it an ideal approach for mixed supervised settings as in the case of the methodology proposed in this thesis.

Previous studies that incorporated cell counts into their model training assumed that these counts were precise. However, to know the exact number of cells, one must count them one by one, which requires almost as much work as obtaining point annotations. On the other hand, in a typical clinical setting, pathologists derive such cell counts approximately using a method known as *eyeballing* [Talat et al., 2023], which corresponds to a preliminary examination of a tissue region and estimating a rough number of cells in this region by sight. This process takes significantly less time than obtaining boundary or point annotations, and the cell count obtained at the end of this process can be considered as the simplest/weakest form of annotation for cell localization and counting tasks. Despite that, *the prediction of cell counts under the supervision of ground truths obtained by eyeballing has not been extensively explored, either as a standalone cell counting task or as an auxiliary task to cell*

localization.

To address this gap, unlike the previous literature, this thesis presents a new mixed-supervised training scheme for a multitask network to exploit ground truths obtained by the eyeballing process, which corresponds to a more realistic setting in digital pathology for acquiring cell count annotations. In this multitask network, each branch is dedicated to a distinct task which will be trained with different supervisory signals determined by the available level of annotations for a training image. In our setting, ground truth cell counts are obtained by eyeballing for every training image, but point annotations are only available for a smaller subset. Since the obtained cell counts are only rough numbers and may be erroneous because of eyeballing, this thesis also introduces a new loss function with a consistency term that penalizes inconsistencies between the cell counts calculated on the predictions of the two branches, in particular to regularize learning for the cell counting task.

1.1 Contributions

The main contributions of this thesis are twofold:

- This is the first proposal of a mixed-supervision approach for digital pathology that utilizes cell counts obtained by eyeballing in a multitask network to simultaneously learn the tasks of cell localization and cell counting.

- A new consistency term is introduced as a regularization technique to enhance task coherence in a multitask network. This term is defined to penalize the differences between the predictions of the two tasks. That is, it is specifically defined to penalize the difference between the number of cells predicted by the cell counting task and the number of cell objects (connected components) in the segmentation map predicted by the cell localization task. Our proposal is different than the previous studies that defined this term between the outputs of multiple cell localization networks, which were designed to predict the same task. As cell localization and cell counting are relevant but more diverse tasks, they are more complementary and the consistency term defined between the

predictions of these tasks is expected to be more regulative.

Testing the proposed methodology on two datasets of hematoxylin-eosin (HE) stained tissue images, our experiments showed that the utilization of cell counts obtained by eyeballing, which can be considered as the weakest form of annotation, in the framework of multitask learning improved the model's performance under the scarcity of stronger annotations.

1.2 Outline

The outline of this thesis is as follows. In Chapter 2, background related to our research is provided under five sections: U-Net network architecture used in this thesis, utilization of point annotations for cell localization, deep learning models employed for cell counting, eyeballing technique, and lastly loss function to ensure consistency in model training. Chapter 3 includes the details of our proposed methodology and network architecture. Chapter 4 provides the information regarding datasets used in our experiments, evaluation metrics and calculations, and comparison algorithms. Chapter 5 contains quantitative and visual results for the proposed methodology and the comparison methods along with the assessment of key findings. Chapter 6 concludes this thesis and discusses future research directions.

Chapter 2

RELATED WORK

This thesis proposes a mixed-supervised learning scheme that utilizes point annotations and cell counts as supervisory signals to train a multitask network. It also introduces a loss with a consistency term to enhance the task coherence. We categorize the related work below according to each aspect.

2.1 U-Net Architecture

In recent years, encoder-decoder networks have become a powerful tool in medical image segmentation, showing effectiveness in handling complex tasks. Among the most widely used encoder-decoder networks for medical applications is U-Net, an architecture that has gained considerable attention. The design of the original U-Net is depicted in Figure 2.1.

The U-Net's encoder, or downsampling component, plays a crucial role in extracting high-level features to represent an image. It functions similarly to a conventional convolutional neural network (CNN), gradually collecting semantic information through successive convolutional layers followed by rectified linear unit (ReLU) activations. Max-pooling is applied to reduce the spatial resolution of the image at each step. With every downsampling operation, the number of feature channels increases, helping to capture more intricate patterns. Since segmentation tasks require a balance of both spatial and semantic information, the decoder section is specifically designed to retain spatial details while processing the semantic content provided by the encoder.

The U-Net's decoder starts at the bottleneck, the final convolutional block in the encoder, and progressively upsamples the feature maps using transposed convolutions. It first receives the semantic features extracted by the encoder and then

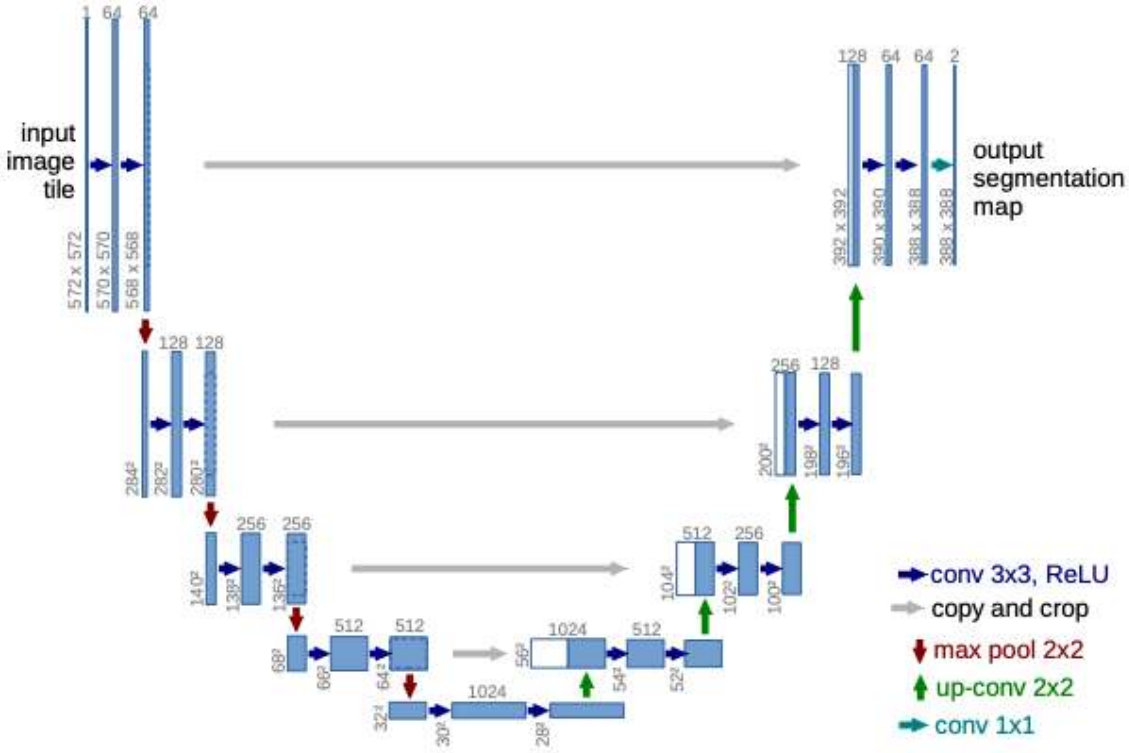


Figure 2.1: The original U-Net architecture. Figure was taken from [Ronneberger et al., 2015].

enriches them with high-resolution feature maps from earlier stages of the network, thanks to skip connections. These connections help retain fine-grained spatial details that are critical for precise segmentation.

Architectures like U-Net offer several advantages for image segmentation tasks. They allow the model to simultaneously capture both global context and detailed local features. Furthermore, U-Net performs well even when training data is limited, making it particularly useful in medical image analysis, where annotated data can be scarce. Lastly, U-Net processes the entire input image at once, preserving the global context through the entire network. This is a significant advantage over patch-based CNN approaches, which divide the image into smaller patches and may lose valuable spatial relationships across the full image.

2.2 Cell Localization with Point Annotations

In the literature, there were two main approaches to make use of point annotations in deep learning-based models. The first was to learn cell locations by training a regression network to estimate density maps generated on the point annotations, while the other was to generate cell segmentation masks from these points. The first group commonly used Gaussians [Qu et al., 2020, Nishimura et al., 2021] and proximity functions [Xie et al., 2018, Pan et al., 2018] to generate the density maps, where the points were considered as peaks and the densities decreased as one moved away from the points, and learned them from original images by training a regression-based dense prediction network. In [Qu et al., 2020], after training a detection model to learn the Gaussian masks, the results were improved by obtaining Voronoi diagrams from the detected regions and combining them with pixel-level k-means clustering. Likewise, in [Chamanzar and Nie, 2020], the repel encoding was used together with the Voronoi diagrams and pixel clustering to enhance cell localization.

The second group generated segmentation maps from the point annotations, commonly by making use of self- and co-training techniques [Yu, 2023, Zhao and Yin, 2021, Lin et al., 2023]. For example, a study in [Yu, 2023] proposed a self-learning approach that shared knowledge between a principal and a collaborator model. The primary model was trained for object detection using the point annotations, while the collaborator model was trained for semantic segmentation under the guidance of the primary model. In [Lin et al., 2023], a self-supervised learning approach was also proposed, but this time, instead of directly using them in training, the point annotations were considered as starting points. The initial step trained a segmentation model to convert the point annotations into rough pixel-level annotations based on Voronoi diagrams and k-means clustering, and a progressive approach was used to obtain more refined labels. In [Yoo et al., 2019], a nucleus segmentation method was developed solely relying on the point annotations. To address the undersegmentation problem and predict detailed nucleus boundaries, this study defined a secondary task and learn it with an additional small network. In [Khalid et al., 2022], the point annotations were used together with the cells' bounding boxes to

generate a segmentation mask. Although additional annotations of the bounding boxes were necessary, it was argued that this level of annotation was still weaker and faster than obtaining the boundary annotations. Our methodology differs from all these studies as it trains a multitask network with different levels of supervision to predict cell locations from the point annotations and makes use of cell counts obtained by eyeballing in this training.

2.3 Deep Learning Models for Cell Counting

One group of studies computed the cell count on the output predicted by a cell segmentation or a detection network. For cell segmentation, the count was calculated as the number of individual objects on the output map predicted by a deep neural network, e.g., by using a fully pyramid network [Hernandez et al., 2018] and a U-Net variant [Dave et al., 2023, Melouk et al., 2023, Chen et al., 2021b, Morelli et al., 2021]. Cell counts were also obtained by counting bounding boxes generated by detection networks, mostly with YOLO models [Tessema et al., 2021, Aldughayfiq et al., 2023], where each box encapsulated a cell object [Zhou et al., 2022]. Additionally, YOLO and U-Net models were used together to calculate the total cell count [Bai et al., 2023]. Another group approached this problem as a regression task, using point annotations to generate density maps. Then, these studies trained a regression network to predict the density maps and calculated cell counts by integrating pixel predictions [Xie et al., 2016, He et al., 2021, Zhu et al., 2021, Guo et al., 2019]. Alternatively, they predicted cell counts by feeding the predicted density maps [Liu et al., 2019a, Liu et al., 2019b, Hagos et al., 2019] or latent space features from an image reconstruction task into a second regression network consisting of convolutional and dense layers [Vununu et al., 2018]. These were either single-task networks that were trained using cell count information as the sole supervisory signal [Xue et al., 2016, Lavitt et al., 2021], or multitask networks that concurrently learned cell count prediction from a shared encoder together with additional tasks of cell segmentation [Huang et al., 2022] or spatial super-resolution [Deng et al., 2023]. Our proposed approach is similar to these multitask networks in that it includes

multiple branches for the tasks of cell localization and cell count prediction. On the other hand, unlike these previous studies, our approach introduces a consistency term in its loss function to ensure coherence in the simultaneous learning of its tasks. Additionally, with the help of this consistency term, it can use cell counts obtained from eyeballing for model supervision, rather than requiring the exact counts.

2.4 Eyeballing Technique

Cell counting is a crucial task in pathology and biomedical research, used for tissue structure analysis, disease diagnosis, and treatment efficacy monitoring. The most accurate method for obtaining this information is one-by-one counting of individual cells. However, this approach is labor-intensive and time-consuming, making it impractical for routine clinical use. Consequently, one-by-one counting is rarely performed in practice. Instead, clinical experts often use a visual assessment technique known as eyeballing. This method relies on the expertise of pathologists to approximate the desired information, such as an index value expressed as the percentage or the count of specific cell types. By avoiding the need for precise cell-by-cell counting, eyeballing offers a less burdensome and more efficient alternative [Akbar et al., 2019, Fulawka and Halon, 2017].

While strict rules for this procedure does not exist, in general, pathologists segment an image into smaller regions to estimate cell density. Afterwards, using the approximates of the cell count in the entire field of view or in a smaller defined area, they extrapolate the result for larger regions. When dealing with large numbers of cells, group counting may be employed, e.g., counting in clusters of five or ten and summing the totals [Hida et al., 2015, Cheng et al., 2020].

Eyeballing has a wide range of applications as a clinical technique. For example, it is used to estimate the ratio of tumor cells in Ki-67 stained tissue images, a marker that serves as a strong indicator of cell proliferation in cancer. This estimation is crucial for calculating the proliferation index and assessing cancer progression [Boden et al., 2021, Dy et al., 2024]. Another notable application is the quantification of tubular atrophy, a recognized method for measuring chronic kidney damage,

performed through eyeballing under light microscopy [Gupta et al., 2022].

Furthermore, gaze-tracking methods have been explored in the literature to reduce the effort required by pathologists and address the shortage of annotated data needed for training deep learning models. For instance, in [Mariam et al., 2022], the authors proposed using gaze tracking of expert pathologists for labeling whole slide images and training object detection models. By comparing results between manually labeled and gaze-labeled data, their proposed methodology demonstrated a significant reduction in annotation time while maintaining satisfactory performance.

Overall, obtaining results through the eyeballing procedure is subject to inter- and intra- observer variability and may be less reproducible. However, expertise in the field of pathology significantly enhances the consistency and reliability of the results [Brunyé et al., 2017]. Additionally, studies in the literature have compared eyeballing results with AI-supported systems and manual counting, often finding strong correlations between them [Talat et al., 2023]. This consistency allows eyeballing to be utilized as ground truth data for model training while significantly reducing annotation efforts.

2.5 Forcing Consistency through a Loss Function

It was used as a regularization technique to force multiple models, trained on the same data, to yield consistent predictions. It helped the models make stable predictions, which usually led to better generalization and performance especially when annotated data was scarce [Peng and Wang, 2021]. The study in [Wu et al., 2022] suggested a multitask network architecture for radiologic image segmentation. It defined a mutual consistency constraint between the probability map of one task and the pseudo-labels of the others to reduce uncertainty for semi-supervised image segmentation. The consistency term can also be defined between the outputs of different tasks. For example, the study in [Kong et al., 2021] proposed a multitask network for endoscopic image classification and segmentation with a cross fusion module, and calculated a consistency loss between the features of the classification task to reduce the misalignment between classification and segmentation

results. Consistency loss functions were also used to train cell segmentation networks. In [Zhao and Yin, 2021], a cell segmentation network was trained with weak supervision, making use of self- and co-training schemes and using a consistency loss to force consensus between a self-trained and two co-trained networks. In [Yu, 2023], cell segmentation was achieved using a principal and a collaborator model that were trained by a self-learning strategy based on knowledge sharing between the models. The knowledge sharing was achieved by defining a consistency loss between the predictions of these models. Different than these previous studies that defined their consistency loss on the results of two segmentation tasks, our proposed method defines it between the prediction of the cell counting task and the number of components in the segmentation map predicted by the cell localization task.

Chapter 3

METHODOLOGY

This thesis proposes a mixed supervised training strategy for a multitask network to concurrently learn the tasks of cell localization and cell counting. This network has one shared encoder and two separate branches, one decoder for cell localization and one fully connected layer for cell counting. The proposed strategy is to train the multitask network when different levels of supervision are available for different instances in a training dataset. It assumes that the training dataset $D = (I_n, Y_n)$, with I_n being the input image and Y_n corresponding to the ground truth annotations available for this image, consists of two subsets $D = D_1 \cup D_2$. In particular, the first subset, $D_1 = \{I_i \,|\, Y_i \in (S \cup C)\}$, contains training images I_i with two types of ground truth annotations: segmentation masks $S_i \in \{0, 1\}$ obtained by dilating point annotations, and cell counts $C_i \in \mathbb{Z}^+$ obtained by counting the annotated points. The second subset, $D_2 = \{I_k \,|\, Y_k \in C\}$, contains training images I_k for which only the cell counts $C_k \in \mathbb{Z}^+$ estimated by expert pathologists using eyeballing is available. Our training strategy updates all network weights when $I_n \in D_1$, whereas it updates the weights of the shared encoder and only the decoder of the cell counting task when $I_n \in D_2$, by backpropagating the joint loss with the proposed consistency term. The overview of the proposed methodology is illustrated in Figure 3.1 and the details are provided in the following subsections.

3.1 Multitask Network Architecture

The multitask network consists of a shared encoder accompanied by a decoder for cell localization and fully connected layers for cell count prediction. The architectures of all networks are illustrated in Figure 3.2. As seen in this figure, the encoder consists of four blocks, each containing two convolutional and one max pooling

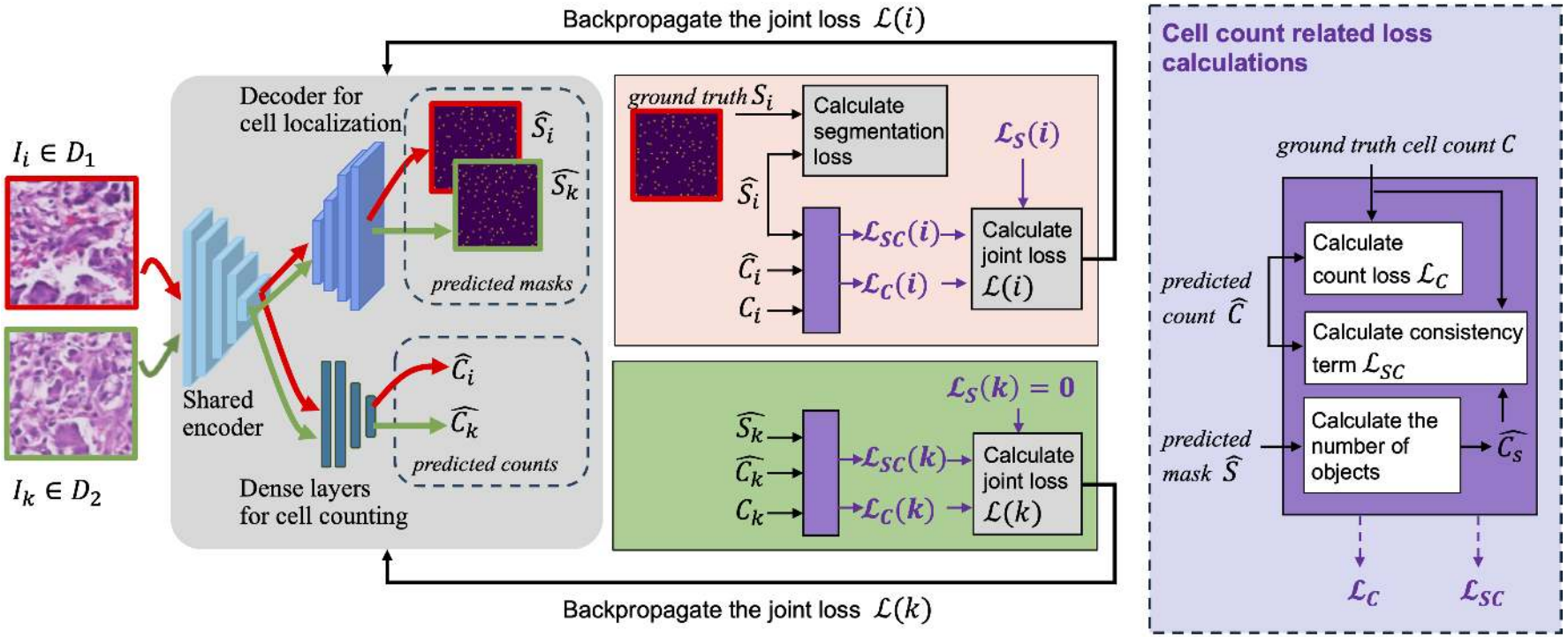


Figure 3.1: Overview of the proposed mixed-supervised strategy to train the multitask network with different levels of supervision. In particular, for a training image $I_i \in D1$, the ground truth mask S_i generated from point annotations and the cell count C_i obtained by counting the annotated points are available, and all loss components are calculated between the ground truths and the predicted values $\hat{S}_i$ and $\hat{C}_i$ (red boxes and arrows). For a training image $I_k \in D_2$, only the cell count C_k obtained by eyeballing is available, and only cell count related loss components are calculated (green boxes and arrows). Note that for I_k, one can also calculate the consistency term $\mathcal{L}_{SC}(k)$ since this calculation uses the ground truth count C_k, its predicted value $\widehat{C_k}$, and the number $\widehat{C_{s_k}}$ of connected components in the predicted mask $\widehat{S_k}$, but not the ground truth mask S_k, which is not available for I_k. This calculation is depicted in the purple box on the right. After calculating the joint loss, relevant network weights are updated through backpropagation.

layer. The convolutions use a 3×3 kernel and a rectified linear unit (ReLu) as the activation function, while the max pooling employs a 2×2 kernel to reduce the spatial dimension. To prevent overfitting, a dropout layer with a rate of 0.2 is added between two consecutive convolutional layers. The bottleneck contains two convolutions and a dropout layer between them. The decoder also consists of four blocks, where in each block, upsampling with a 2×2 kernel is applied to recover the original input size, and its output is concatenated with the features of the corresponding encoder block before being fed into a convolution with a 3×3 kernel. At the end, the sigmoid function is used to obtain binary pixel labels. In the branch of cell count prediction, global average pooling is first applied to the features obtained from the shared encoder, followed by two dense layers with 64 and 32 hidden units, respectively, both using a ReLu as the activation function. At the end, the linear function is used to obtain cell counts, as it corresponds to a regression task.

3.2 Settings for Separate Branches

The first branch is for dense prediction to locate cells on an image. It is supervised by ground truth masks that contain rough locations of the cells. These masks are generated by dilating point annotations provided by pathologists. For any given ground truth mask S_n, the pixel value is 1 if a pixel belongs to the cell area after dilation, and 0 otherwise. For an image $I_n \in D_1$, the loss $\mathcal{L}_S(n)$ of this branch is the binary cross entropy, which is defined between the pixel values in S_n and those in the segmentation map $\widehat{S_n}$ predicted by the network. It is worth noting that this loss can be calculated for training images $I_n \in D_n$ with point annotations, and the weights for the decoder of this branch are updated. For the other training images $I_n \in D_2$, the loss of this branch $\mathcal{L}_S(n)$ is set to 0, since there are no point annotations to generate the ground truth masks, and thus, there is no update of the decoder's weights for these training images.

The second branch corresponds to a regression problem for cell count prediction. It is supervised by the ground truth count C_n obtained by counting the annotated

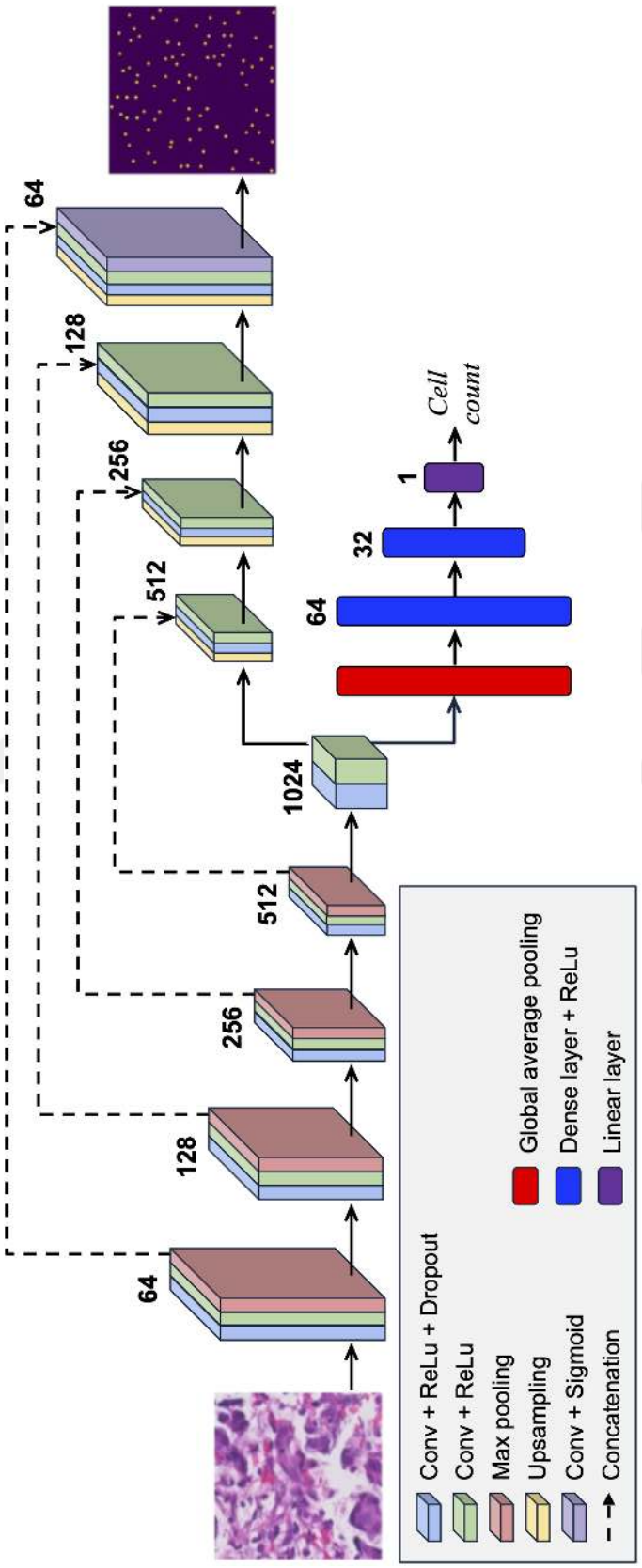


Figure 3.2: Architecture of the multitask network used in our proposed model. Each colored box represents a certain operation in a block. The number of feature maps used in each block is also indicated on the top of this block.

points for the images in the first set, $I_n \in D_1$, and by eyeballing for those in the second set, $I_n \in D_2$. For an image I_n, the loss $\mathcal{L}_C(n)$ of this branch is defined as

$$\mathcal{L}_C(n) = \frac{|C_n - \widehat{C_n}|}{C_n} \tag{3.1}$$

where $\widehat{C_n}$ is the cell count predicted by the network. It is an absolute error between the ground truth and the predicted cell counts, normalized by the ground truth. This normalization ensures that the loss is comparable for all images, regardless of the number of cells they contain. This loss is defined for all training images in $D = D_1 \cup D_2$ because it does not require knowing the point annotations. Hence, the weights of the dense layers of this branch are updated for all training images.

3.3 Mixed-Supervised Network Training with Joint Loss

The two tasks of cell localization and cell count prediction concurrently learned by optimizing the following joint loss function $\mathcal{L}(n)$ with the proposed consistency term.

$$\mathcal{L}(n) = \begin{cases} \mathcal{L}_S(n) & +\alpha\ \mathcal{L}_C(n) + \beta\ \mathcal{L}_T(n), \quad \text{if } I_n \in D_1 \\ 0 & +\alpha\ \mathcal{L}_C(n) + \beta\ \mathcal{L}_T(n), \quad \text{if } I_n \in D_2 \end{cases} \tag{3.2}$$

where $\mathcal{L}_T(n)$ is the consistency term for the image I_n, and α and β are the relative contributions of the cell count loss and the consistency term to the joint loss function, respectively. The consistency term for the image I_n is defined as

$$\mathcal{L}_T(n) = \frac{|\widehat{C_n} - \widehat{C_{s_n}}|}{C_n} \tag{3.3}$$

where C_n is the ground truth, $\widehat{C_n}$ is the cell count predicted by the cell counting (second) branch of the network, and $\widehat{C_{s_n}}$ is the number of cell objects (connected components) in the mask $\widehat{S_n}$ predicted by the cell localization (first) branch of the network. This term can also be calculated for all training images in $D = D_1 \cup D_2$ because it does not use the ground truth mask S_n, but the predicted mask $\widehat{S_n}$. As the dense layers of the cell count prediction task are optimized using the weakest form of supervision, which may contain error due to the nature of the eyeballing process, this consistency term is used to update the weights of the dense layers of this task

to regularize its predictions. Additionally, since the cell localization branch may produce noisy masks in early epochs, the consistency term is used after a warm-up period, it is not used in the first 25 epochs. The weights of the shared encoder are updated by backpropagating all loss terms available for a given training image.

All networks are constructed and trained in the Python programming language using TensorFlow. The Adam optimizer is used to train the networks. The learning rate and exponential decay parameters are set to 0.0001 and 0.9, respectively, and the batch size is 1. The training continues for a maximum of 200 epochs, and the network weights that give the best loss value on validation images are selected.

Chapter 4

EXPERIMENTS

4.1 Datasets

We conducted our experiments on two datasets of HE stained tissue images. The first is the in-house serous carcinoma dataset, comprising 109 images with a resolution of 1536×1536 pixels. These images were cropped from multiple whole slide images of various subjects. The data collection was conducted in accordance with the tenets of the Declaration of Helsinki and approved by Koç University Institutional Review Board (protocol number: 2021.336.IRB2.066). Point annotations of these images were performed by a team of pathologists. After months of obtaining these point annotations, a pathologist was shown each image and asked to estimate the cell count by eyeballing. Although both annotations were available for all images, when training a network, we used one annotation or the other, depending on how we partitioned the training dataset into D_1 and D_2. When $I_n \in D_1$, we generated a segmentation map by dilating the point annotations with a disk structuring element. The disk radius was selected as 3 considering the average cell size in the images of this dataset. Then, we used the generated segmentation map as the ground truth mask S_n and the number of the annotated points as the ground truth count C_n. When $I_n \in D_2$, we used the cell count obtained by eyeballing as the only ground truth C_n.

To understand the effectiveness of the proposed training strategy against data scarcity, we performed our experiments with different partition percentages, each time putting p percent of the training data into the first dataset D_1, and the rest into the weakly annotated dataset D_2. For the in-house serous carcinoma dataset, we used three-fold cross validation. In each trial, we used all images in two folds for training; we used approximately 80 percent of the images to learn the network

weights and 20 percent as validation data to select the best network weights. We used the other fold as a test set to evaluate the model performance. We calculated the performance metrics on the test images using their cell counts obtained by counting the point annotations, rather than those obtained by eyeballing, since this would gave more reliable results.

The second was a publicly available dataset, MoNuSeg, that comprised HE stained tissues of various organs taken from multiple patients [Kumar et al., 2017]. Originally, it contained 37 training and 14 test images. In the experiments, we kept this training and test splits, only using 20 percent of the training images as validation data. To ensure compatibility of an input image with the depth requirements of our multitask network's architecture, we cropped four non-overlapping patches, with a resolution of 496×496 pixels, from each original image whose resolution was 1000×1000. Likewise, we conducted multiple experiments by putting p percent of the training data into D_1 and the rest into D_2, for different p values. Originally, the MoNuSeg dataset had the maps of cell boundary annotations for all images. However, to simulate a weaker supervision with point annotations and eyeballing, we did not use the boundary annotations. Instead, we calculated the centroids of the cells and used them as the point annotations. The ground truth masks were then generated by dilating the point annotations. In addition, the same team of pathologists were asked to provide the cell counts of each image in the MoNuSeg dataset by eyeballing.

For each of these datasets, the numbers of images and cells in the training, validation, and test sets are given in Table 4.1. Note that we conducted three-fold cross validation for the in-house dataset, while using the original training and test splits for the MoNuSeg dataset. Also note that the serous carcinoma dataset contains images with a significantly higher number of cells, which makes point or boundary annotations much more challenging. For such images, the weakest form of providing cell counts by eyeballing is much more efficient. To emphasize the difficulty of obtaining point or boundary annotations, Figure 4.2 and Figure 4.1 shows patches cropped from an original images of the MoNuSeg and serous carcinoma datasets, respectively. Note that the patches shown in these figures were cropped for better

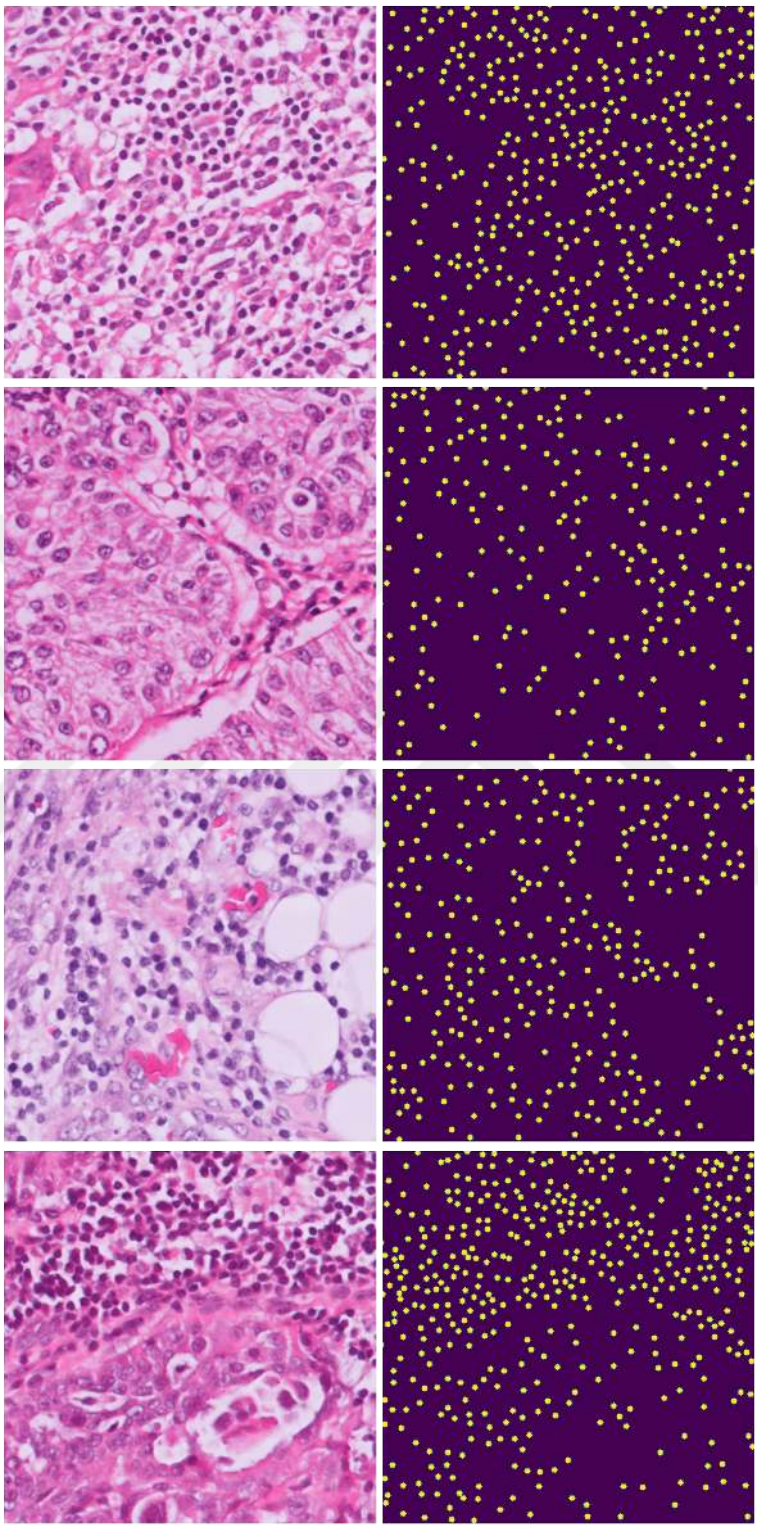
illustration. For example, in Figure 4.1, each patch has a resolution of 384×384 pixels and represents only 1/16th of the original image in the serous carcinoma dataset, which has a resolution of 1536×1536 pixels.

For the serous carcinoma dataset, the reference ground truth maps were obtained using the dot annotations placed at the cell centers, rather than delineating the exact boundaries of cell, as the latter is a more labor-intensive task. However, to train the proposed pipeline, two key targets were required: the cell masks S and the number of cells C_n. The value of C_n was simply defined as the number of cells in the input patch obtained by eyeballing procedure. Pseudo-segmentations of cells were generated from the dot annotations using below equation. In this equation, *S(i,j)* represents the pixel intensity at location *(i,j)* in the pseudo-segmentation image S, d denotes the Euclidean distance between pixel *(i,j)* and any cell dot annotation, and r is a threshold distance. In our experiments the threshold was empirically set to 3 pixels for serous carcinoma dataset and 5 pixels for MoNuSeg dataset to ensure that the pseudo-segmentations of different cells did not overlap.

$$S(i,j) = \begin{cases} 1 & \text{if } d < r \\ 0 & \text{otherwise} \end{cases} \tag{4.1}$$

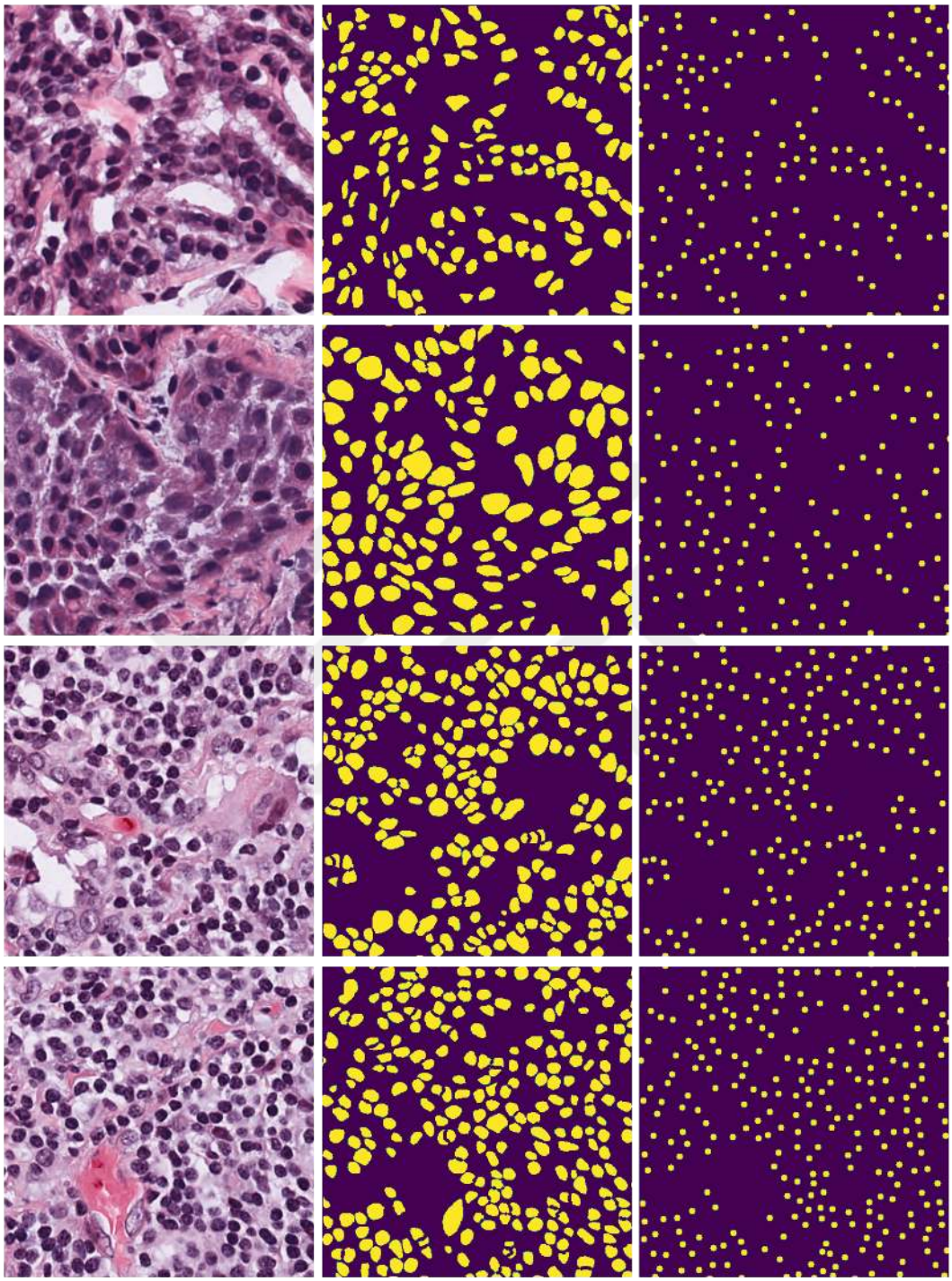
4.2 Evaluations

We evaluated the cell localization and cell counting tasks quantitatively on the images of the test set/fold. For cell localization, we found true positive cells in each predicted map, and calculated the object-level precision, recall, and F1-score metrics. These metrics were calculated for each image separately, and then averaged over the test set/fold. Since ground truth labels consist of point annotations, a predicted object was deemed a true positive only if it matched with a corresponding ground truth center. In the serous carcinoma dataset, the dense population and rich variation in small cell structures occasionally led to the misclassification of correctly generated cell regions as false negatives. To mitigate this, predicted cell objects were also considered true positives if the Euclidean distance between the predicted object center and the ground truth center was less than 10 pixels, excluding cases with multiple



(a) (b)

Figure 4.1: Visual examples selected from our in-house serous carcinoma dataset. (a) Patches cropped from the original images. (b) Point annotations generated by dilating the dot annotations provided by expert pathologists. The patches were cropped for better illustration.



(a) (b) (c)

Figure 4.2: Visual examples selected from the MoNuSeg dataset. (a) Patches cropped from the original images. (b) Boundary annotations of each cell object in the corresponding image. Note that boundary annotations are provided with the original version of the MoNuSeg dataset. (c) Point annotations generated by finding the centroids of the cells for which the boundary annotations are known. The patches were cropped for better illustration.

Table 4.1: For the two datasets used in the experiments, the numbers of images and cells in the training (Tr), validation (Val), and test (Ts) sets. Note that we used three-fold cross validation for the serous carcinoma dataset, and provided these numbers for each trial separately.

Dataset	Number of images			Number of cells		
	Tr	Val	Ts	Tr	Val	Ts
Serous (trial 1)	59	14	36	65711	15895	43025
Serous (trial 2)	58	14	37	68251	17497	38883
Serous (trial 3)	58	15	36	66509	15399	42723
MoNuSeg	120	28	56	16794	7180	6619

matches. Afterwards precision metric is calculated by dividing true positive object counts to total predicted cell count, recall metric by dividing true positive object count to total ground truth count, then F1-score was calculated as harmonic mean of precision and recall metrics. Same approach was adopted while evaluating model predictions on the MoNuSeg dataset either. Even though the MoNuSeg dataset provides boundary level annotations, our models trained with ground truth maps obtained by dilating the cell center information during training, simulating the cell detection task thus during test phase generated maps provides cell locations without complete boundary information e.g. smaller and distorted cell objects. Therefore, object level segmentation metrics such as intersection over union (IoU) were not preferred for evaluation. Note that, we used the predictions as they were, without any postprocessing, for the serous carcinoma dataset, while we eliminated regions smaller than 35 pixels for the MoNuSeg dataset. We experimentally selected this small region threshold and used for our method and all comparison algorithms.

For cell count prediction, we calculated the relative difference $|C_n - \widehat{C_n}|/C_n$ for each test image I_n, between its ground truth C_n and the predicted cell count $\widehat{C_n}$, and averaged them over the entire test set/fold. We first computed this metric by taking the value predicted by the cell counting branch as $\widehat{C_n}$. Then we also

computed it by taking the number of cell objects in the mask $\widehat{S_n}$ predicted by the cell localization branch as the predicted cell count. We will refer to them as $\mathrm{RD}_{\mathrm{Count}}$ and $\mathrm{RD}_{\mathrm{Loc}}$, respectively. Note that in these calculations, we used the numbers of annotated points as the ground truths of test images, but not the counts obtaining by eyeballing.

4.3 Comparison Algorithms

We compared our proposed training strategy with two models: ConCORDe-Net [Hagos et al., 2019] and SSRNet [Deng et al., 2023]. The first model, ConCORDe-Net, was also designed for learning the two tasks of cell detection and counting. However, different than our proposal, it first detected cells using an encoder-decoder network, and then used the generated cell detection mask as the input to the cell counting module. To have a fair comparison, we used the same encoder-decoder network architecture and the same loss for the cell detection task. On the other hand, we implemented the cell count prediction module as suggested by [Hagos et al., 2019] and trained this module with respect to their loss function, which calculates a similarity measure by inverting mean absolute error between the ground truth and predicted cell counts over the batch, then subtracting it from 1 to obtain final loss value between 0 and 1. Note that this loss function did not contain any consistency term.

The ConCORDe-Net model was a deep learning-based framework tailored for cell detection and classification in multiplex immunohistochemically stained whole slide images. It was adapted from Inception-v3, which was developed to address cell detection and counting tasks. The model integrated a cell count loss function with the conventional Dice overlap loss as a part of its objective function, thus aimed to enable the network to detect closely located and weakly stained cells more effectively. The cell count loss regularized the network to learn weak gradient boundaries, facilitating the separation of faintly stained cells from background artifacts. The architecture incorporated a multi-stage convolutional neural network for accurate cell classification. Since we focused on cell detection/localization, we did not include its

cell classification module to our experiments.

The second comparison model was SSRNet [Deng et al., 2023] that also learned cell count prediction in a multitask learning framework. For that, it constructed a multitask network with the tasks of cell segmentation and cell count predictions. This study generated a cell segmentation map by reconstructing an input image with a spatial-based super-resolution module, and regressed the cell count from the features extracted with the upsampling module of this reconstruction. This suggested model was promoted as a non-point based counting method. Thus, in our experiments, instead of using point annotations to generate cell segmentation masks, we trained their models to reconstruct the input images, providing both cell segmentation maps and cell count predictions as outputs.

The SSRNet model was based on an encoder-decoder architecture, performing two tasks: cell counting and cell distribution generation. The encoder utilized a simplified version of the VGG-16 architecture, which reduced its size to one-third of the original network while maintaining its capacity to extract cell features faster. To enhance the spatial detail, SSRNet incorporated a spatial-based super-resolution fast upsampling module. This module performs a one-step upsampling of feature maps, enlarging them by 32 times without introducing information loss or distortion, thus enriching the spatial content. Additionally, an optimized multitask loss function was proposed to coordinate the simultaneous training of counting and cell distribution tasks.

Chapter 5

RESULTS AND DISCUSSION

We trained our proposed approach (*MixedSupervision*) on a training set $D = D_1 \cup D_2$ with mixed supervisory signals. Specifically, D_1 contains point annotations, with the number of annotated points serving as the cell counts, and D_2 contains only the cell counts obtained by eyeballing. In our experiments, we used p percent of the training images in D_1 and the rest in D_2, where p determines the percentage of data requiring labor-intensive annotation. The primary goal of this work is to design a model that reduces the annotation effort required for learning cell localization and counting tasks. To investigate this, we first analyzed the effects of different p percentages on training performance. Tables 5.1a and 6.1b show the results for the serous carcinoma and MoNuSeg datasets, respectively. Since network weight initialization affects the final model, we repeated each experiment three times for our model, as well as for the comparison algorithms and all ablation studies. The values in Table 5.1 represent the average test fold results from nine runs (three folds, with three runs per fold) and Table 5.2 reports the average test set results from three runs (as there was a single test set for the MoNuSeg dataset).

Table 5.1b and 5.2b report comparative results for the single-task counterparts. We trained two single-task networks. The first one learned cell localization from the masks generated using the point annotations available in D_1, which contained only p percent of the entire training set. After training, we calculated the object-level F1-score on the prediction masks of the test images, and also reported $\mathrm{RD}_{\mathrm{Loc}}$, which was calculated as the relative difference between the ground truth and the number of cell objects in the prediction mask. The second single-task network learned cell counting on the entire training set $D = D_1 \cup D_2$. In training, the annotated point counts were used as ground truths for images in D_1, and the cell counts obtained by eyeballing were used as ground truths for images in D_2. Tables 5.1 and 5.2 show that

the multitask networks outperformed their single-task counterparts, especially in cell counting. This improvement can be attributed to the shared encoder, which helps learn more representative features when cell localization is used as a complementary task.

Next, we analyzed the effects of using the consistency term in the joint loss function. Tables 5.1c and 5.2c also report the results obtained without using the consistency term. They revealed that adding a consistency term to the combined loss function improved both cell localization and cell counting. This indicates the effectiveness of incorporating an additional consistency term in the cell counting branch to regularize the learning of the cell count prediction task, as well as to enhance the learning of the shared encoder when point annotations were unavailable in the training set. The results highlighted the benefits of using this loss function and the multitask design under data scarcity.

All these results suggested that good performance can be achieved even when only 25 percent of training images had point annotations, reducing the annotation effort by 75 percent. We compared our approach with the ConCORDe-Net algorithm, evaluating ConCORDe-Net also for $p = \{100, 75, 50, 33, 25\}$ percent. Since the SSRNet algorithm is a non-point annotation method, we trained it using the complete training set.

5.1 Comparisons

The quantitative results on the test sets are given in Table 5.3 and Table 5.4 for the serous carcinoma and the MoNuSeg datasets, respectively. The results reveal that our proposed approach leads to better object-level scores and lower regression-based cell counting errors. Comparing our approach with ConCORDe-Net, which had a cascaded multitask architecture and a different loss function for penalizing cell count predictions, the results indicated the structuring multitask networks with a shared encoder provides more robust and representative features, increasing effectiveness of both tasks. Importance of including a shared encoder and learning common features across multiple tasks becomes more understandable when a portion of ground truth

Table 5.1: For the serous carcinoma dataset, the test fold metrics were obtained by (a) the proposed *MixedSupervision* model, (b) its single-task counterparts, and (c) where the consistency term was not used in the joint loss functon. Results are the averages of the nine runs from three folds and, their standard deviations.

(a)

$p\%$	*MixedSupervision*		
	F1-score	RD_{Loc}	RD_{Count}
100	84.45 ± 1.84	0.14 ± 0.04	0.19 ± 0.05
75	83.87 ± 1.80	0.13 ± 0.03	0.20 ± 0.06
50	82.77 ± 1.82	0.14 ± 0.02	0.22 ± 0.02
33	82.05 ± 2.10	0.15 ± 0.02	0.26 ± 0.11
25	81.45 ± 1.95	0.15 ± 0.03	0.28 ± 0.05

(b)

$p\%$	Single-task cell localization		Single-task cell counting
	F1-score	RD_{Loc}	RD_{Count}
100	83.34 ± 1.66	0.13 ± 0.04	0.24 ± 0.02
75	82.15 ± 1.40	0.14 ± 0.03	0.24 ± 0.03
50	82.02 ± 1.83	0.14 ± 0.02	0.26 ± 0.02
33	81.68 ± 1.42	0.15 ± 0.03	0.27 ± 0.05
25	80.08 ± 1.63	0.16 ± 0.03	0.29 ± 0.02

(c)

$p\%$	*MixedSupervision* – w/o consistency		
	F1-score	RD_{Loc}	RD_{Count}
100	83.69 ± 1.82	0.14 ± 0.03	0.24 ± 0.11
75	82.69 ± 1.90	0.15 ± 0.04	0.35 ± 0.20
50	82.16 ± 2.16	0.15 ± 0.03	0.41 ± 0.21
33	81.30 ± 2.21	0.14 ± 0.02	0.41 ± 0.18
25	80.99 ± 1.39	0.16 ± 0.03	0.44 ± 0.12

Table 5.2: For the MoNuSeg dataset, the test fold metrics were obtained by (a) the proposed *MixedSupervision* model, (b) its single-task counterparts, and (c) where the consistency term was not used in the joint loss function. Results are the averages of the three runs and their standard deviations.

(a)

$p\%$	*MixedSupervision*		
	F1-score	RD_{Loc}	RD_{Count}
100	80.86 ± 0.44	0.07 ± 0.01	0.14 ± 0.02
75	79.10 ± 0.55	0.07 ± 0.03	0.15 ± 0.03
50	77.68 ± 0.28	0.09 ± 0.03	0.19 ± 0.03
33	77.32 ± 0.81	0.08 ± 0.02	0.19 ± 0.03
25	76.41 ± 0.11	0.09 ± 0.03	0.18 ± 0.01

(b)

$p\%$	Single-task cell localization		Single-task cell counting
	F1-score	RD_{Loc}	RD_{Count}
100	78.57 ± 0.53	0.08 ± 0.01	0.28 ± 0.02
75	77.60 ± 1.07	0.09 ± 0.02	0.31 ± 0.02
50	75.17 ± 1.85	0.12 ± 0.02	0.32 ± 0.04
33	75.09 ± 1.64	0.09 ± 0.02	0.35 ± 0.01
25	73.34 ± 1.16	0.11 ± 0.04	0.47 ± 0.02

(c)

$p\%$	*MixedSupervision* – w/o consistency		
	F1-score	RD_{Loc}	RD_{Count}
100	79.57 ± 1.07	0.07 ± 0.01	0.22 ± 0.12
75	77.86 ± 2.23	0.08 ± 0.01	0.22 ± 0.07
50	75.14 ± 1.56	0.10 ± 0.02	0.33 ± 0.12
33	75.63 ± 1.01	0.09 ± 0.01	0.32 ± 0.05
25	73.93 ± 1.04	0.10 ± 0.02	0.50 ± 0.06

Table 5.3: For the serous carcinoma dataset, test set/fold results obtained by the proposed *MixedSupervision* approach and the comparison algorithms when p percent training data had point annotations. These are averages of 9 runs from 3 folds.

p%	Model	Serous carcinoma dataset		
		F1-score	RD_{Loc}	RD_{Count}
100	*MixedSupervision*	84.45 ± 1.84	0.14 ± 0.04	0.19 ± 0.05
	ConCORDe-Net	82.29 ± 2.60	0.16 ± 0.03	0.26 ± 0.15
	SSRNet	55.60 ± 2.04	0.29 ± 0.01	0.34 ± 0.04
75	*MixedSupervision*	83.87 ± 1.80	0.13 ± 0.03	0.20 ± 0.06
	ConCORDe-Net	79.92 ± 2.15	0.18 ± 0.02	0.30 ± 0.02
50	*MixedSupervision*	82.77 ± 1.82	0.14 ± 0.02	0.22 ± 0.02
	ConCORDe-Net	77.59 ± 1.63	0.19 ± 0.05	0.31 ± 0.03
33	*MixedSupervision*	82.05 ± 2.10	0.15 ± 0.02	0.26 ± 0.11
	ConCORDe-Net	76.21 ± 2.03	0.21 ± 0.03	0.37 ± 0.05
25	*MixedSupervision*	81.45 ± 1.95	0.15 ± 0.03	0.28 ± 0.05
	ConCORDe-Net	74.18 ± 1.75	0.27 ± 0.04	0.40 ± 0.12

labels for supervising one of the tasks is scarce. To demonstrate such a scenario, we compared our approach with ConCORDe-Net training both algorithms by using reduced amount of the ground truth point annotations from the train set, thus hindering gradient flow of weight update mechanism for cell segmentation task. Note that results for all of our models are obtained with included consistency terms, however discarding this term our design still performs better than cascaded network architecture 5.1c and 5.2c.

Compared with the SSRNet algorithm, the proposed models yielded drastically better F1-scores and lower cell counting errors on both datasets. Despite SSRNet had a similar multitask architecture with shared encoder, integrated image reconstruction module aimed for non-point based cell distribution performs poorly on images taken from HE datasets, where inter-cellular tissue that serves as background

Table 5.4: For the MoNuSeg dataset, test set/fold results obtained by the proposed *MixedSupervision* approach and the comparison algorithms when p percent training data had point annotations. These are averages of 3 runs as there was a single test set.

p%	Model	MoNuSeg dataset		
		F1-score	RD_{Loc}	RD_{Count}
100	*MixedSupervision*	80.86 ± 0.44	0.07 ± 0.01	0.14 ± 0.02
	ConCORDe-Net	74.37 ± 0.70	0.11 ± 0.02	0.19 ± 0.02
	SSRNet	50.98 ± 1.29	0.21 ± 0.01	0.27 ± 0.03
75	*MixedSupervision*	79.10 ± 0.55	0.07 ± 0.03	0.15 ± 0.03
	ConCORDe-Net	72.55 ± 0.93	0.12 ± 0.02	0.20 ± 0.02
50	*MixedSupervision*	77.68 ± 0.28	0.09 ± 0.03	0.19 ± 0.03
	ConCORDe-Net	70.94 ± 1.12	0.14 ± 0.05	0.23 ± 0.02
33	*MixedSupervision*	77.32 ± 0.81	0.08 ± 0.02	0.19 ± 0.03
	ConCORDe-Net	67.24 ± 0.73	0.17 ± 0.02	0.30 ± 0.04
25	*MixedSupervision*	76.41 ± 0.11	0.09 ± 0.03	0.18 ± 0.01
	ConCORDe-Net	66.17 ± 1.73	0.18 ± 0.02	0.38 ± 0.04

is diverse and complex. Due to this complexity, reconstructed images fail to serve as accurate segmentation maps with under segmented regions where multiple cell objects appear as one large area. Visual results obtained for both datasets highlight this issue by representing the center of a segmented area with a single green dot instead of marking each cell center with one green dot which indicates correct detection. Although most of the calculated centers are inaccurate and fall onto background not single cellular regions, yet correct predictions exist.

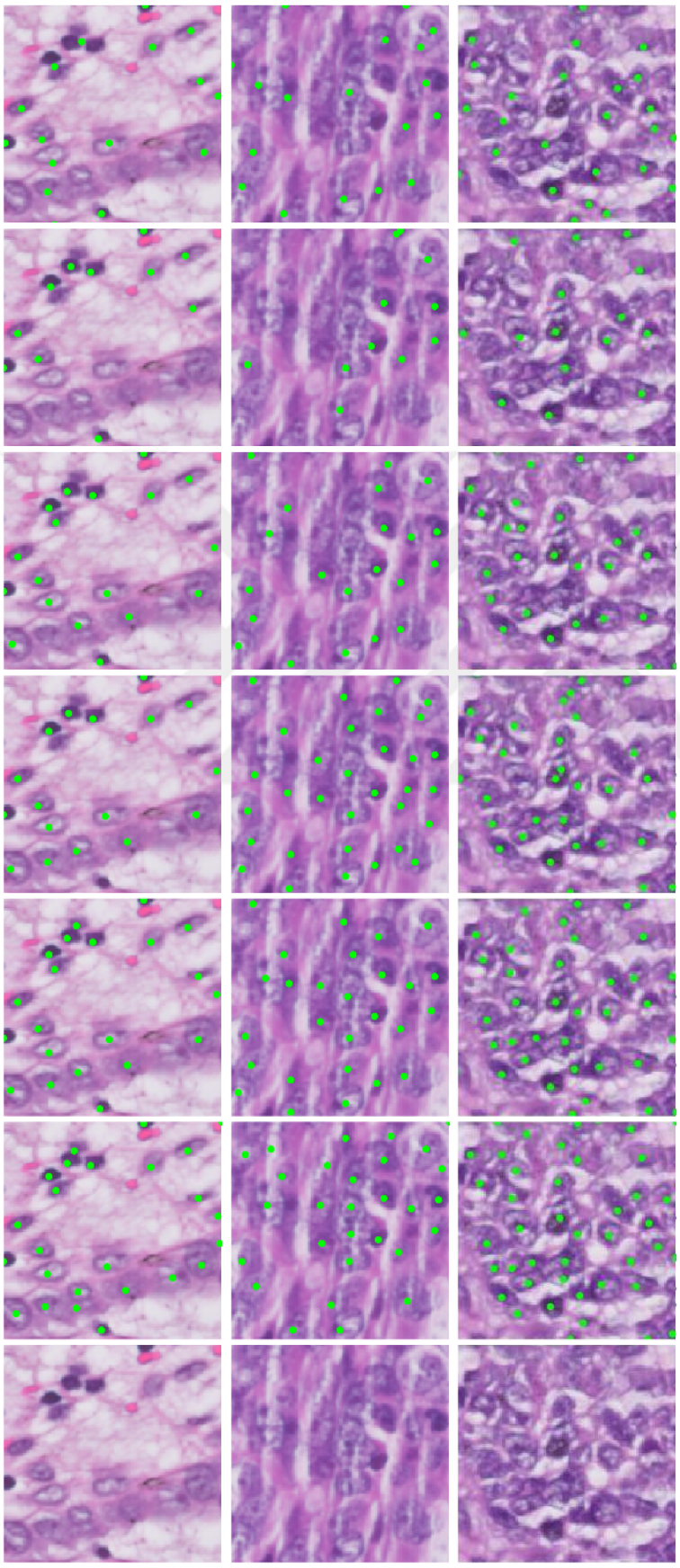
We have conducted extensive ablation studies on both datasets by limiting ground truth dot annotation amounts that supervise the cell detection branch to observe the effect of data scarcity on overall detection and cell counting performance. As expected, reducing the ground truth label amount had a negative effect on cell

detection metric as well as cell count predictions. However, despite this collapse on metrics, multitask network architecture performed better than the it's counterpart single-task architecture. This improvement can be attributed to deploying the cell counting branch as an auxiliary task, thus even though the cell detection branch did not receive any gradient response during training, weights of the shared encoder is updated by cell counting loss and, if used together, consistency term, tolerating data scarcity.

Additionally, each algorithm used on previously mentioned ablation study run including and discarding consistency term during training of the model. Aim of this study is to observe the effect of this term on penalizing cell count predictions whether it enhances coherence between both tasks with guidance of generated segmentation maps. Results shared on Table 5.1a and Table 6.1b indicated significant improvements on cell number predictions after adding consistency term to initial loss function with decay of cell counting error regardless the percentage of point annotations used for supervising cell segmentation branch.

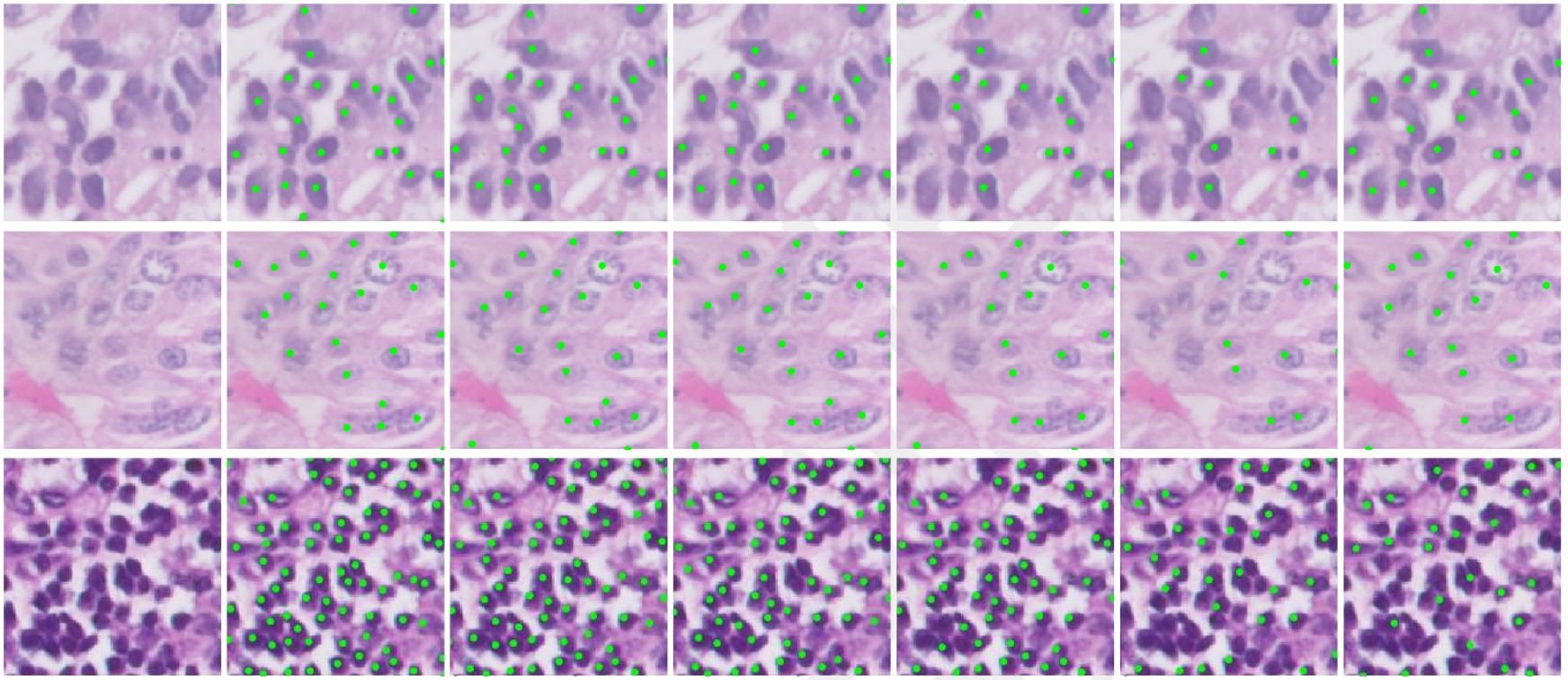
These quantitative results align with the visual observations obtained from the serous carcinoma dataset (Figures 5.1 and 5.2) and the MoNuSeg dataset (Figures 5.3 and 5.4). As seen on the regions corresponding to green colored dots our proposed approaches detected target cell regions more successfully. Moreover, false positives cases such as predicting inter-cellular tissue as a cell object are less frequent than compared algorithms. Visual results obtained from two different datasets point out that our approach is robust enough to locate various cell types with different shape and sizes from diverse tissue images represented with distinct colors and magnification. There are failure cases where our approaches are also unable to successfully detect cell objects, however for most of these contexts, comparison models mistaken with SSRNet perform the worst, thus overall performance of our models prevails the compared algorithms.

Our approach achieved better object-level F1-scores and lower relative difference errors. The relative improvement was even higher as p decreased. This might suggest that our proposed loss function, with the inclusion of the consistency term, facilitated the learning of a more robust and representative shared encoder, thereby



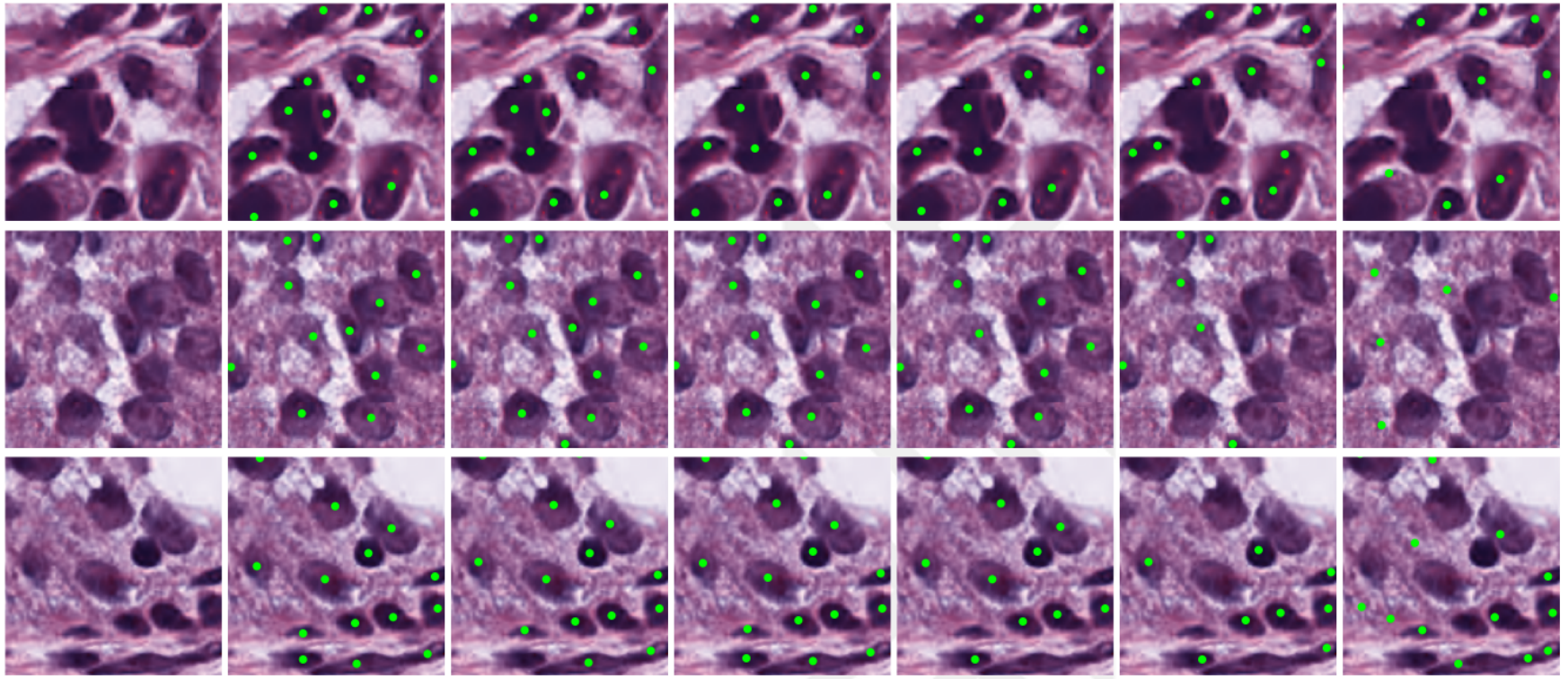
(a) (b) (c) (d) (e) (f) (g)

Figure 5.1: Visual results on example test set images. (a) Patches cropped from the original images. All images selected from the our in-house serous carcinoma dataset. The patches were cropped for better illustration. (b) Point annotations in the ground truths. (c) *MixedSupervision* when $p = 100$, (d) *MixedSupervision* when $p = 25$, (e) ConCORDe-Net [Hagos et al., 2019] when $p = 100$, (f) ConCORDe-Net when $p = 25$, and (g) SSRNet [Deng et al., 2023].



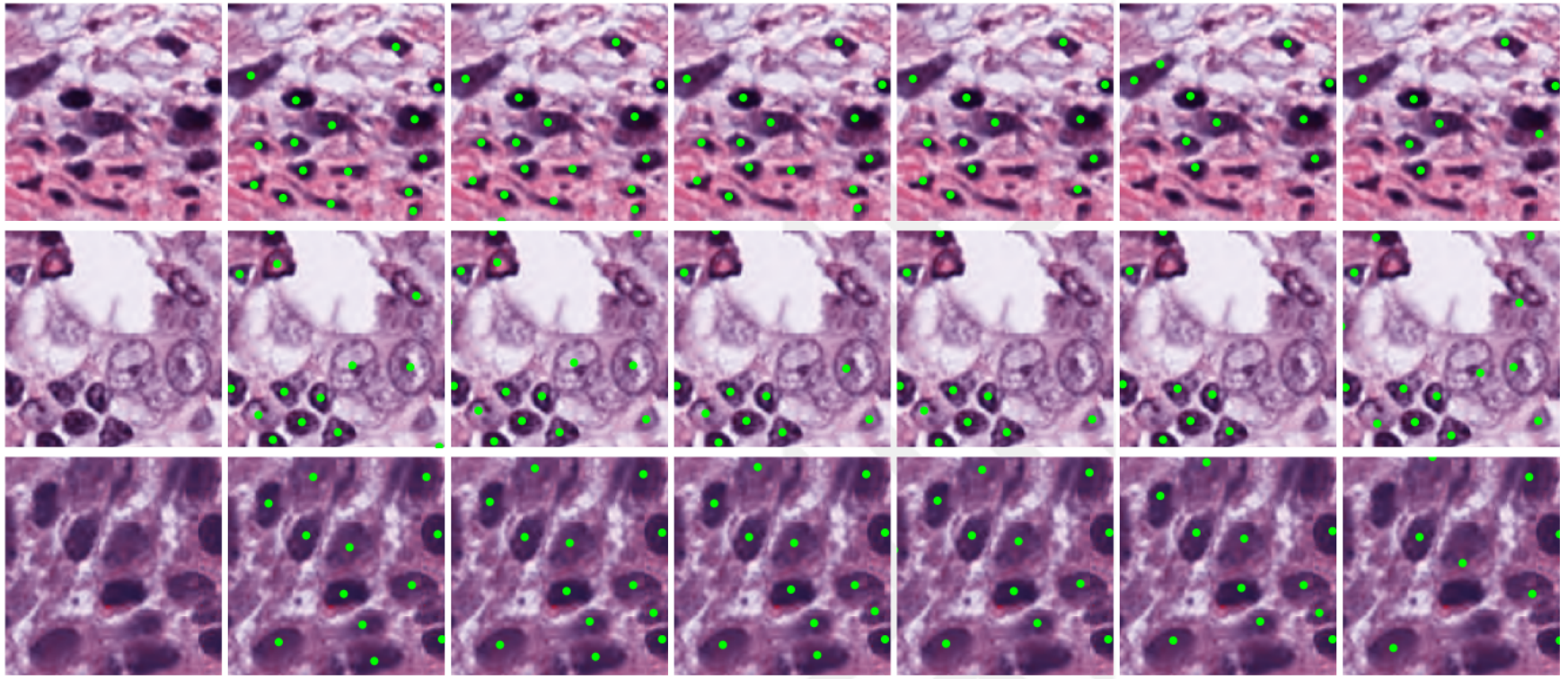
(a) (b) (c) (d) (e) (f) (g)

Figure 5.2: Visual results on example test set images. (a) Patches cropped from the original images. All images selected from the our in-house serous carcinoma dataset. The patches were cropped for better illustration. (b) Point annotations in the ground truths. (c) *MixedSupervision* when $p = 100$, (d) *MixedSupervision* when $p = 25$, (e) ConCORDe-Net [Hagos et al., 2019] when $p = 100$, (f) ConCORDe-Net when $p = 25$, and (g) SSRNet [Deng et al., 2023].



(a) (b) (c) (d) (e) (f) (g)

Figure 5.3: Visual results on example test set images. (a) Patches cropped from the original images. All images selected from the MoNuSeg dataset. The patches were cropped for better illustration. (b) Point annotations in the ground truths. (c) *MixedSupervision* when $p = 100$, (d) *MixedSupervision* when $p = 25$, (e) ConCORDe-Net [Hagos et al., 2019] when $p = 100$, (f) ConCORDe-Net when $p = 25$, and (g) SSRNet [Deng et al., 2023].



(a) (b) (c) (d) (e) (f) (g)

Figure 5.4: Visual results on more example test set images. (a) Patches cropped from the original images. All images selected from the MoNuSeg dataset. The patches were cropped for better illustration. (b) Point annotations in the ground truths. (c) *MixedSupervision* when $p = 100$, (d) *MixedSupervision* when $p = 25$, (e) ConCORDe-Net [Hagos et al., 2019] when $p = 100$, (f) ConCORDe-Net when $p = 25$, and (g) SSRNet [Deng et al., 2023].

enhancing the effectiveness of both tasks, even with the scarcity of strongly annotated data. Provided figures presents visual results of this improvement. Note that these are smaller patches cropped from the original images for better illustration. As seen in these figures, although ConCORDe-Net correctly identified most of the cells when all training data had point annotations (columns e), the accuracy decreased when $p = 25$ percent (columns f). On the other hand, the proposed *MixedSupervision* approach led to better results when $p = 100$ percent, as well as the accuracy decrease was smaller when p decreases to 25 percent (columns c and columns d). The second comparison model was SSRNet, which also employs a multitask architecture with a shared encoder. However, the results indicated that its integrated image reconstruction module, designed for non-point-based cell distribution, performed poorly on HE images, where the inter-cellular tissue serves as a diverse and complex background. This complexity caused reconstructed images to fail in providing accurate segmentation maps, leading to undersegmented cells (columns g).

5.2 Analyzes of Loss Weights

Our proposed approach utilize two external parameters each of which weights the two loss functions combined for penalizing cell count predictions. One parameter α is used as weighting factor on the effect of cell count loss and other is β used weighting the contribution of consistency term to overall cell counting loss. We have experimented with two sets of candidate loss weights to analyze effect of each term by changing one parameter while keeping other constant is fixed. Figure 5.5 and 5.6 represent the change on evaluation metrics as a result of using different weighting schemes on cell count loss, where $\alpha \in \{0.01, 0.03, 0.05, 0.07, 0.1, 0.3, 0.5, 0.7, 1, 2\}$. Results reveal that selecting larger values for α worsen the performance of cell segmentation branch, while selecting a very small values causes model to miss optimal condition. Similarly, increase in α constant had benefit for cell counting branch up to a certain value. This observation can be attributed to including consistency term in which effect of it depends on cell segmentation branch, thus increasing contribution from cell count branch to shared encoder via larger loss constant, ac-

curacy of segmentation branch is worsen hence the cell count predictions. Figures 5.7 and 5.8 demonstrate the results of similar experiment on β parameter, where $\beta \in \{0.005, 0.007, 0.01, 0.05, 0.07, 0.1, 0.5, 0.7, 1, 2\}$ fixing the α parameter. It is shown that choosing values for β between 0 and 1 results in better segmentation performance and lower cell counting error. Note that loss constant for segmentation/detection loss is fixed to value of one and all experiments were conducted by employing two different training scenarios, considering provided outcomes, hyperparameter selection for auxiliary task loss weights is crucial to obtain betterment on results. Otherwise, due to dependency between different tasks over shared encoder and the consistency term model performance degrades.

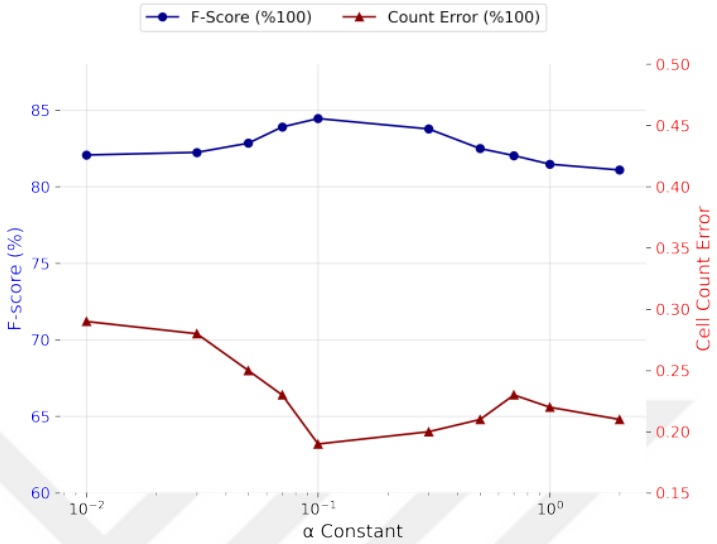


Figure 5.5: Visualization of changes on cell segmentation and cell count metrics as response to change in weight of cell count loss, α, when $p = 100$.

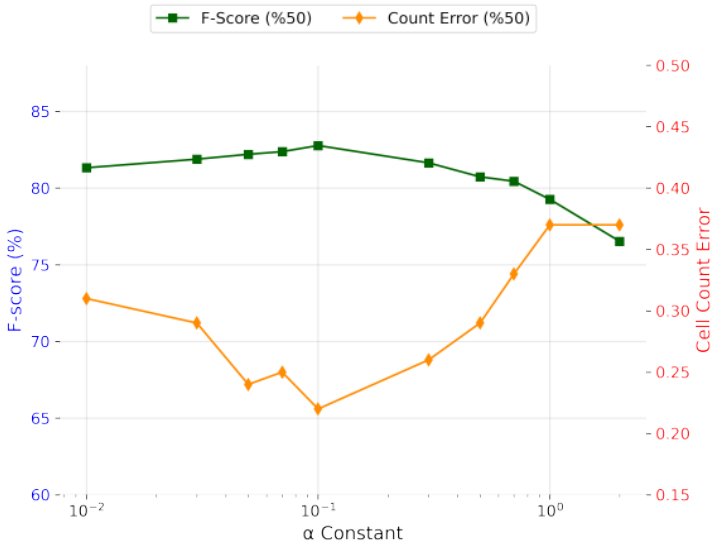


Figure 5.6: Visualization of changes on cell segmentation and cell count metrics as response to change in weight of cell count loss, α, when $p = 50$.

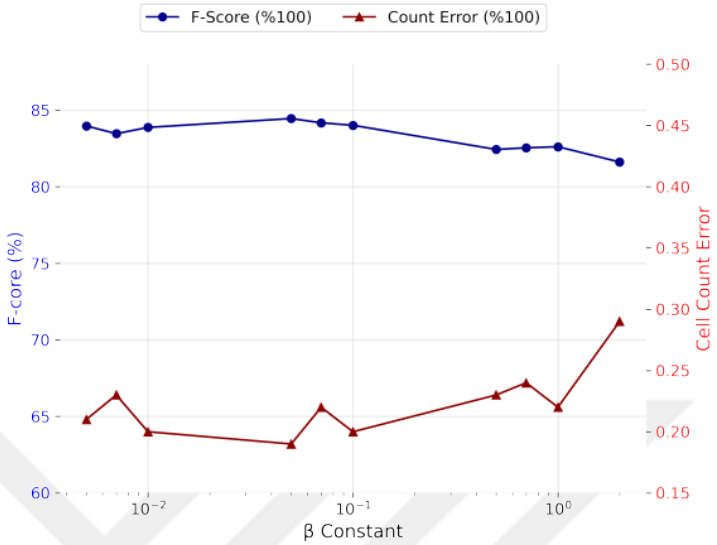


Figure 5.7: Visualization of changes on cell segmentation and cell count metrics as response to change in weight of consistency loss, β, when $p = 100$.

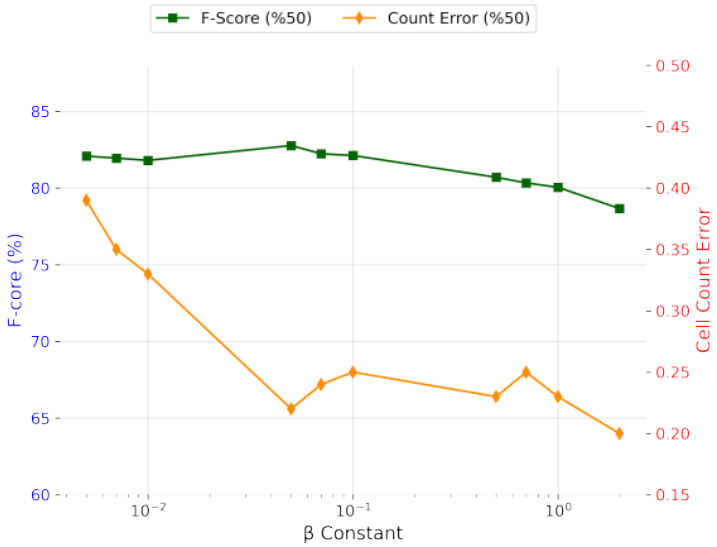


Figure 5.8: Visualization of changes on cell segmentation and cell count metrics as response to change in weight of consistency loss, β, when $p = 50$.

Chapter 6

CONCLUSION

This thesis introduced a multitask network design employing a mixed-supervised training approach to simultaneously learn the tasks of cell counting and cell localization. The proposed *MixedSupervision* approach relied on using two types of ground truth labels: point annotations to guide cell localization, and eyeballing-derived cell counts, which represented the weakest and least labor-intensive form of annotation, to supervise cell counting. To ensure consistency between the predictions of these two tasks, the *MixedSupervision* approach introduced a consistency term in the definition of its loss function. We tested our proposed approach on two datasets of hematoxylin-eosin stained tissue images. Our experiments demonstrated that defining cell count prediction as an auxiliary task within a multitask network and using the eyeball-derived cell counts as an additional supervisory signal in its training effectively regularized the cell localization task, even when stronger point annotations were limited. As a result, it led to better performance compared to its counterparts.

This study is the first proposal of integrating eyeballing-derived ground truths into network training. It highlights the potential to increase the availability and amount of supervisory signals for training networks in digital pathology, without relying on resource-intensive annotations. In future work, the potential of mixed-supervision approaches can be further explored and applied to a diverse range of tasks in digital pathology, unlocking new opportunities for innovation and advancement in the field. For instance, tumor diagnosis or classification, which "partially" rely on conventional eyeballing techniques, can be significantly improved by training networks with a mixed-supervision approach. This innovative strategy has the potential to revolutionize the field of digital pathology, enabling more accurate and efficient diagnoses/prognosis.

We trained the models used in this thesis from scratch, learning the weights for

cell segmentation/detection and cell counting tasks simultaneously. Alternatively, multitask models can be trained by employing pre-trained single-task networks as a starting point, with a concept similar to transfer learning.

Table 6.1 demonstrates the results of this alternative approach. Instead of training our mixed-supervision approach from scratch, we initialized the model using the weights of single-task networks learned in previous experiments and added a cell counting branch. In our previous experiments, the consistency term was incorporated after the 25th epoch to address under-segmented results of the segmentation/detection branch, in its first epochs, which complicated the cell counting process. However, in this new scenario, since the pre-trained segmentation/detection network had already converged, the consistency term was applied starting from the first epoch.

Additionally, while all models in the previous experiments were trained for up to 200 epochs, in this new scenario, the models were trained for a maximum of 100 epochs. All other hyperparameters were kept the same. Results reported in Table 6.1 indicate that this approach achieved satisfactory performance, surpassing all comparison algorithms on both datasets for equivalent training data usage percentages. However, they are slightly lower than the results obtained by the proposed multitask model, trained from scratch. This might be attributed to the effectiveness of simultaneously updating the shared encoder from the start of training. In contrast, the effect of the weight updates from the cell counting branch was limited due to the use of an already converged model.

Table 6.1: Experimental results of *MixedSupervision* by employing the weights of the pre-trained single task segmentation/detection networks. Results of the proposed *MixedSupervision* model trained with the weights of the corresponding single task networks on (a) the serous carcinoma dataset and (b) the MoNuSeg dataset. Results are averages of multiple runs (nine runs from three folds for the serous carcinoma dataset, and three runs for the MoNuSeg dataset).

(a)

$p\%$	F1-score	$\mathrm{RD}_{\mathrm{Loc}}$	$\mathrm{RD}_{\mathrm{Count}}$
100	83.91 ± 0.24	0.08 ± 0.02	0.22 ± 0.03
50	82.48 ± 0.34	0.10 ± 0.02	0.26 ± 0.05
25	81.02 ± 0.22	0.09 ± 0.04	0.30 ± 0.04

(b)

$p\%$	F1-score	$\mathrm{RD}_{\mathrm{Loc}}$	$\mathrm{RD}_{\mathrm{Count}}$
100	79.98 ± 0.20	0.08 ± 0.02	0.18 ± 0.02
50	76.17 ± 0.18	0.09 ± 0.03	0.22 ± 0.04
25	74.91 ± 0.23	0.08 ± 0.02	0.25 ± 0.03

6.1 Future Work

This thesis has demonstrated the feasibility and benefits of a mixed-supervision approach for cell detection and counting. There are several future research directions and potential applications that can further expand its impact. These are summarized as follows.

- The mixed-supervision framework can be applied to a broader range of tasks in digital pathology, such as tumor segmentation, classification, and grading. These tasks often partially rely on the traditional eyeballing technique, making them ideal candidates for the integration of weakly supervised signals.

- The proposed framework can be extended to integrate multi-modal data, such as combining pathological images with radiological scans or genomic data. This would allow the network to learn complementary features and improve its generalization across different diagnostic tasks.

- Further exploration of advanced multitask architectures, such as attention mechanisms and transformer-based models, could improve the robustness of predictions. This would be particularly useful for scenarios involving heterogeneous datasets or varied annotation quality.

- The proposed *MixedSupervision* approach can be tented across larger and more diverse datasets, including different tissue types and staining methods. This would validate its generalizability and effectiveness in various pathological contexts.

- Hybrid approaches that combine weak supervision (eyeballing) with semi-supervised learning methods can be explored. By leveraging both labeled and unlabeled data, these approaches could further enhance the training process while reducing reliance on labor-intensive manual annotations.

- Active learning strategies can be incorporated to selectively refine annotations

by asking pathologists to provide stronger labels only for ambiguous cases. This would optimize the balance between weak and strong supervision.

The *MixedSupervision* approach proposed in this thesis represents a significant step forward in reducing the dependency on resource-intensive annotations for training deep learning models in digital pathology. By leveraging the strengths of both weak and strong supervision, it opens up new possibilities for improving accuracy, scalability, and efficiency in various pathological tasks.

BIBLIOGRAPHY

[Akbar et al., 2019] Akbar, S. et al. (2019). Automated and manual quantification of tumour cellularity in digital slides for tumour burden assessment. *Sci. Rep.*, 9.

[Aldughayfiq et al., 2023] Aldughayfiq, B., Ashfaq, F., Jhanjhi, N. Z., and Humayun, M. (2023). YOLOv5-FPN: a robust framework for multi-sized cell counting in fluorescence images. *Diagnostics*, 13.

[Alom et al., 2020] Alom, Z., Aspiras, T., Taha, T., Bowen, T., and Asari, V. (2020). MitosisNet: End-to-end mitotic cell detection by multi-task learning. *IEEE Access*, PP:1–1.

[Bai et al., 2023] Bai, H., Wang, X., Guan, Y., Gao, Q., and Han, Z. (2023). Blood cell counting based on U-Net++ and YOLOv5. *Optoelectron. Lett.*, 19:370–376.

[Boden et al., 2021] Boden, A. et al. (2021). The human-in-the-loop: An evaluation of pathologists' interaction with ai in clinical practice. *Histopathology*, 79.

[Brunyé et al., 2017] Brunyé, T., Mercan, E., Weaver, D., and Elmore, J. (2017). Accuracy is in the eyes of the pathologist: The visual interpretive process and diagnostic accuracy with digital whole slide images. *J. Biomed. Inform.*, 66.

[Chamanzar and Nie, 2020] Chamanzar, A. and Nie, Y. (2020). Weakly supervised multi-task learning for cell detection and segmentation. In *IEEE Int. Symp. Biomed. Imag.*, pages 513–516.

[Chen et al., 2021a] Chen, J. et al. (2021a). TransUNet: Transformers make strong encoders for medical image segmentation. *arXiv:2102.04306*.

[Chen et al., 2021b] Chen, Y., Liang, D., Bai, X., Xu, Y., and Yang, X. (2021b). Cell localization and counting using direction field map. *IEEE J. Biomed. Health Inf.*, 26:359–368.

[Cheng et al., 2020] Cheng, J., Abel, J., Balis, U., McClintock, D., and Pantanowitz, L. (2020). Challenges in the development, deployment regulation of artificial intelligence (ai) in anatomical pathology. *Am. J. Pathol.*, 191.

[Chow et al., 2020] Chow, Z. L. et al. (2020). Counting mitoses with digital pathology in breast phyllodes tumors. *Arch. Pathol. Lab. Med.*, 144.

[Dave et al., 2023] Dave, P. et al. (2023). MIMO U-Net: efficient cell segmentation and counting in microscopy image sequences. In *SPIE Med. Imag.*, page 30.

[Deng et al., 2023] Deng, L., Zhou, Q., Wang, S., and Zhang, Y. (2023). SSRNet: A deep learning network via spatial-based super-resolution reconstruction for cell counting and segmentation. *Adv. Intell. Syst.*, 5.

[Dy et al., 2024] Dy, A. et al. (2024). Ai improves accuracy, agreement and efficiency of pathologists for ki67 assessments in breast cancer. *Sci. Rep.*, 14.

[Frei et al., 2023] Frei, A. et al. (2023). Pathologist computer-aided diagnostic scoring of tumor cell fraction: A Swiss national study. *Mod. Pathol.*, 36:100335.

[Fujita and Han, 2020] Fujita, S. and Han, X. (2020). Cell detection and segmentation in microscopy images with improved mask R-CNN. In *ACCV Workshops*.

[Fulawka and Halon, 2017] Fulawka, L. and Halon, A. (2017). Ki-67 evaluation in breast cancer: The daily diagnostic practice. *Indian J. Pathol. Microbiol.*, 60:177.

[Graham et al., 2019] Graham, S. et al. (2019). Hover-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Med. Image Anal.*, 58:101563.

[Graham et al., 2023] Graham, S. et al. (2023). One model is all you need: Multi-task learning enables simultaneous histology image segmentation and classification. *Med. Image Anal.*, 83:102685.

[Greenwald et al., 2022] Greenwald, N. et al. (2022). Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. *Nat. Biotechnol.*, 40:555–565.

[Guo et al., 2019] Guo, Y., Stein, J., Wu, G., and Krishnamurthy, A. (2019). SAU-Net: a universal deep network for cell counting. In *ACM Int. Conf. Bioinf. Comput. Biol. Health Inf.*, volume 2019, pages 299–306.

[Gupta et al., 2022] Gupta, K., Maitra, D., and Gowrishankar, S. (2022). Whole slide imaging vs eyeballing: The future in quantification of tubular atrophy in routine clinical practice. *Indian J. Nephrol.*, 32:151–155.

[Hagos et al., 2019] Hagos, Y. B., Narayanan, P. L., Akarca, A. U., Marafioti, T., and Yuan, Y. (2019). ConCORDe-Net: cell count regularized convolutional neural network for cell detection in multiplex immunohistochemistry images. In *Int. Conf. Med. Image Comput. Comput. Assist. Intervent.*

[He et al., 2021] He, S., Minn, K., Solnica-Krezel, L., Anastasio, M., and Li, H. (2021). Deeply-supervised density regression for automatic cell counting in microscopy images. *Med. Image Anal.*, 68:101892.

[Hernandez et al., 2018] Hernandez, C., Sultan, M., and Pande, V. (2018). Using deep learning for segmentation and counting within microscopy data. *arXiv:1802.10548*.

[Hida et al., 2015] Hida, A. et al. (2015). Visual assessment of ki67 using a 5-grade scale (eye-5) is easy and practical to classify breast cancer subtypes with high reproducibility. *J. Clin. Pathol.*, 68.

[Huang et al., 2022] Huang, L., Huang, L., and Yang, M. (2022). MSRCN: multi-task learning network for cell segmentation and regression counting. In *Int. Symp. Artif. Intell. Med. Sci.*, pages 486–490.

[Khalid et al., 2022] Khalid, N. et al. (2022). Point2Mask: A weakly supervised approach for cell segmentation using point annotation". In *Med. Image Understanding Anal.*, pages 139–153.

[Khalid et al., 2023] Khalid, N. et al. (2023). PACE: point annotation-based cell segmentation for efficient microscopic image analysis. In *Int. Conf. Artif. Neural Networks*.

[Kong et al., 2021] Kong, Z. et al. (2021). Multi-task classification and segmentation for explicable capsule endoscopy diagnostics. *Front. Mol. Biosci.*, 8.

[Kumar et al., 2017] Kumar, N., Verma, R., Sharma, S., Bhargava, S., Vahadane, A., and Sethi, A. (2017). A dataset and a technique for generalized nuclear segmentation for computational pathology. *IEEE Trans. Med. Imag.*, 36(7):1550–1560.

[Lavitt et al., 2021] Lavitt, F., Rijlaarsdam, D., Linden, D., Weglarz-Tomczak, E., and Tomczak, J. (2021). Deep learning and transfer learning for automatic cell counting in microscope images of human cancer cell lines. *Appl. Sci.*, 11:4912.

[Lin et al., 2023] Lin, Y. et al. (2023). Nuclei segmentation with point annotations from pathology images via self-supervised learning and co-training. *Med. Image Anal.*, 89:102933.

[Liu et al., 2019a] Liu, Q., Junker, A., Murakami, K., and Hu, P. (2019a). Automated counting of cancer cells by ensembling deep features. *Cells*, 8:1019.

[Liu et al., 2019b] Liu, Q., Junker, A., Murakami, K., and Hu, P. (2019b). A novel convolutional regression network for cell counting. In *IEEE Int. Conf. Bioinf. Comput. Biol.*, pages 44–49.

[Mariam et al., 2022] Mariam, K. et al. (2022). On smart gaze based annotation of histopathology images for training of deep convolutional neural networks. *IEEE J. Biomed. Health Inform.*, PP:1–1.

[Melouk et al., 2023] Melouk, O., Klingner, A., Elmasry, R. M., and Salem, M. A.-M. (2023). Auto-encoders for detection and counting of live/dead cells. In *Int. Conf. Intell. Comput. Inf. Syst.*, pages 45–51.

[Miao et al., 2021] Miao, R., Toth, R., Zhou, Y., Madabhushi, A., and Janowczyk, A. (2021). Quick annotator: An open-source digital pathology based rapid image annotation tool. *J. Pathol. Clin. Res.*, 7.

[Morelli et al., 2021] Morelli, R. et al. (2021). Automating cell counting in fluorescent microscopy through deep learning with c-ResUnet. *Sci. Rep.*, 11:22920.

[Nishimura et al., 2021] Nishimura, K., Wang, C., Watanabe, K., Ker, D. F. E., and Bise, R. (2021). Weakly supervised cell instance segmentation under various conditions. *Med. Image Anal.*, 73:102182.

[Oh et al., 2022] Oh, H.-J., Lee, K., and Jeong, W.-K. (2022). Scribble-supervised cell segmentation using multiscale contrastive regularization. In *IEEE Int. Symp. Biomed. Imag.*, pages 1–5.

[Pan et al., 2018] Pan, X. et al. (2018). Cell detection in pathology and microscopy images with multi-scale fully convolutional neural networks. *World Wide Web*, 21.

[Peng and Wang, 2021] Peng, J. and Wang, Y. (2021). Medical image segmentation with limited supervision: A review of deep network models. *IEEE Access*, 9:36827–36851.

[Qu et al., 2020] Qu, H. et al. (2020). Weakly supervised deep nuclei segmentation using partial points annotation in histopathology images. *IEEE Trans. Med. Imag.*, 39(11):3655–3666.

[Ronneberger et al., 2015] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Int. Conf. Med. Image Comput. Comput. Assist. Intervent.*, pages 234–241.

[Talat et al., 2023] Talat, Z., Shams, M., Ahmad, Z., Chundriger, Q., Ahmed, A., and Jaffar, N. (2023). Ki-67 quantification in breast cancer by digital imaging AI software and its concordance with manual method. *J. Coll. Physicians Surg. Pak.*, 33:544–547.

[Tessema et al., 2021] Tessema, A., Mohammed, M., Simegn, G., and Kwa, T. (2021). Quantitative analysis of blood cells from microscopic images using convolutional neural network. *Med. Biol. Eng. Comput.*, 59.

[Vununu et al., 2018] Vununu, C., Kwon, O.-H., Kwon, K.-R., Lee, S.-H., and Kang, K.-W. (2018). A deep feature learning scheme for counting the cells in microscopy data. In *IEEE Int. Conf. Electron. Commun. Eng.*, pages 22–26.

[Wu et al., 2022] Wu, Y. et al. (2022). Mutual consistency learning for semi-supervised medical image segmentation. *Med. Image Anal.*, 81:102530.

[Xie et al., 2016] Xie, W., Noble, J., and Zisserman, A. (2016). Microscopy cell counting and detection with fully convolutional regression networks. *Comp. Meth. Biomech. Biomed. Eng. Imag. Vis.*, pages 1–10.

[Xie et al., 2018] Xie, Y., Xing, F., Shi, X., Kong, X., Su, H., and Yang, L. (2018). Efficient and robust cell detection: A structured regression approach. *Med. Image Anal.*, 44:245–254.

[Xue et al., 2016] Xue, Y., Ray, N., Hugh, J., and Bigras, G. (2016). Cell counting by regression using convolutional neural network. In *ECCV Workshops*, volume 9913, pages 274–290.

[Yoo et al., 2019] Yoo, I., Yoo, D., and Paeng, K. (2019). PseudoEdgeNet: nuclei

segmentation only with point annotations. In *Int. Conf. Med. Image Comput. Comput. Assist. Intervent.*

[Yu, 2023] Yu, J. (2023). Point-supervised single-cell segmentation via collaborative knowledge sharing. *IEEE Trans. Med. Imag.*, 42(12):3884–3894.

[Zhao and Yin, 2021] Zhao, T. and Yin, Z. (2021). Weakly supervised cell segmentation by point annotation. *IEEE Trans. Med. Imag.*, 40(10):2736–2747.

[Zhou et al., 2022] Zhou, Y. et al. (2022). ErythroidCounter: An automatic pipeline for erythroid cell detection, identification and counting based on deep learning. *Multimed. Tools Appl.*, 81.

[Zhu et al., 2021] Zhu, Y., Chen, Z., Zheng, Y., Zhang, Q., and Wang, X. (2021). Real-time cell counting in unlabeled microscopy images. In *Int. Conf. Comp. Vis. Workshops*, pages 694–703.