Single-Shot Object Detection and Supervised Image Segmentation for Analysing Cell Images Obtained by Immunohistochemistry

Gustavo Martins L. da Costa*, Anna P. C. Rodrigues *, Gabriel Barbosa da Fonseca*
Zenilton K. G. do Patrocínio Jr*, Giovanna Ribeiro Souto*, Silvio Jamil F. Guimarães*

*Laboratory of Image and Multimedia Data Science
Pontifical Catholic University of Minas Gerais, Brazil, 31980–110
gumartinslopes@gmail.com, annapugac@gmail.com, {gabrielfonseca,zenilon,grsourto,sjamil}@pucminas.br

***Abstract*—Analyzing cell images and identifying them correctly is a fundamental task in the immunohistochemical exam. In this paper we propose a novel method to segment FoxP3+ Regulatory T cells (Treg) images automatically, in order to assist healthcare professionals in the task of identifying and counting potentially cancerous cells. The proposed method relies on combining an object detection network, which is tailor-made for microscopy images, with a marker-based image segmentation method to produce the final segmentation, while requiring only a 50x50 training patch to do so. Our pipeline consists on predicting the location of the cells, applying morphological operations on the prediction weights to transform them into markers, and finally using the segmentation method iDISF to generate high quality segmentations. We also propose a new FoxP3+ Treg cells dataset containing 10 high resolution images, with a qualitative and quantitative analysis of our segmentation methods for this dataset.**

I. Introduction

Immunohistochemistry is a widely used laboratory technique in the field of pathology and biomedical research to analyze and identify specific proteins in tissue samples. For instance, one widely studied protein is the FoxP3+, which is one of the main markers distinguishing regulatory T-cells from T-cells [1]. Many studies have been made on the detection and analysis of FoxP3+ presence and ratio on tissue samples, to assess its impact on anti-tumor immune response [2], on the development of oral squamous cells carcinoma [3], pancreatic ductal adenocarcinoma [4], and hepatocarcinoma [5]. These studies suggest that the level of FoxP3+ may be used as a prognostic factor and can bring interesting clinical implications [2]. Thus, correctly detecting and counting these cells is of great interest.

Even though immunohistochemistry is a standard process for staining and highlighting specific types of cells, several difficulties are faced on the analysis of the produced images. The most common strategy for analysing images obtained from this process is based on a visual inspection made by manual counting performed by the pathologist using a conventional microscope which evaluates and quantifies of positively immunostained cells. Possible interpretation and counting errors can occur due to wear and tear during the analysis process, which significantly affects decision-making about the analyzed tissue.

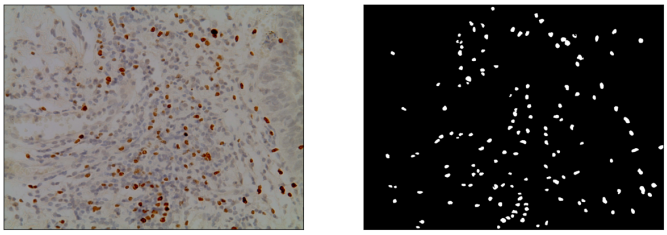

Fig. 1. **Ground truth example of an image obtained by immunohistochemistry.** The image on the left is a component of the original FoxP3+ cells dataset. The image on the right side demonstrates the outcome after the process of obtaining a ground truth annotation.

The time for analysing these types of databases can be considerably long, taking into account the experience of the professional who performs the exam. This type of examination also requires extensive training for professionals to correctly identify the cells to be counted. Considering the points presented above, automating the identification of these cells will lead to significant gains in efficiency and quality in immunohistochemical examinations. In Figure 1, we illustrate a sample of an image obtained by immunohistochemistry, and a possible annotation for it producing its groundtruth.

In order to improve the reliability of the process, avoiding subject evaluation, and to decrease the time for analysing images, several methods have been proposed in literature [6], [7]. However, most of them are complicated and hand-crafted, and sometimes require previously annotated datasets in order to be tuned. In this work, we propose a strategy for automatically segmenting immunostained Regulatory T-cells (Treg), based on a single-shot learning and interactive graph-based image segmentation. We also produce a new dataset containing images with immunostained Regulatory T-cells and their groundtruth. The former is used to identify cell markers to be used in the supervised image segmentation, and the latter for producing robust cell delineation.

The main contributions of this work are threefold: (i) we

propose a method for automatically producing segmentations for FoxP3+ cells datasets that can be trained with only a single cell image; (ii) we propose a new FoxP3+ cells dataset, containing images sampled from mouth tissues; and (iii) a baseline for the proposed dataset, with a quantitative and qualitative analysis.

The remainder of this article is organized as follows: in Section II we describe briefly some basic concepts for a full understanding of the entire process; in Section III we give an overview of the image dataset used; a detailed view of the entire pipeline is provided in Section IV; experimental setup and results are presented in Section V; finally, in Section VI is conclusions and future work are drawn.

II. Basic Concepts

In this section, we present iDISF and ultimate erosion which can be considered as, in conjuction of a single-shot network, two very critical parts of our proposed strategy for segmenting T-cells.

A. Interactive Dynamic and Iterative Spanning Forest

Image segmentation is one of the main tasks in image processing and computer vision applications. It consists of dividing an image into meaningful regions or components, with the aim of isolating objects of interest from the rest of the scene, referred to as the background. Such technique is widely used in computer vision to extract the most important information from an image for a given context.

The Interactive Dynamic and Iterative Spanning Forest (iDISF) is an interactive segmentation method derived from the DISF superpixel computation method [8]. The DISF method is able to compute very accurate superpixel delineation, but it does not relate the superpixels to any object of interest. On iDISF, we have the introduction of a human given prior, which indicates the location of the object of interest, and with this information, we are able to get an object segmentation with the same delineation quality given by the DISF method.

Let I be an image, S_O be the set of seeds given by the user representing the object of interest and S_B be set of background seeds, also given by the user. The first step of iDISF is to expand the S_B set by oversampling background seeds over the image, by using a grid sampling strategy, as in [9]. We denote the expanded background seeds by S_B^e. The final segmentation $Seg^I(S_O, S_B^e)$ is given by iteratively computing the optimum-path forest [10] rooted on the given set of seeds, filtering the least relevant seeds at each step. The seed removal criteria are detailed in [8]. A sample of a cell image segmentation produced with iDISF is illustrated in Figure 2.

To compute a segmentation using iDISF, the user must give an initial set of object seeds, indicating the location of the objects of interest. These initial seeds guide the computation of the optimum path forests and the removal of badly sampled background seeds, and thus have a great importance to the method. Even though the inclusion of the object seeds help iDISF to compute robust segmentations, they usually come from user interactions.

Fig. 2. **iDISF object manual marking process.** The image on the left shows a fragment of a original dataset image, in which there are blue markers on top of FoxP3$^+$ cells. The image on the right shows the result of the iDISF segmentation given such markers.

B. Ultimate Erosion

The ultimate erosion is a mathematical morphology strategy made, for example, by iterative erosion of the image until all objects vanish. Here, the idea is to identify markers (or seeds) for the cells, thus the primary goal of efficiently reducing a specified object within a binary image to a compact representation composed of only a few pixels. Given a grayscale image, the result of an ultimate erosion over it can be defined as the set of connected components. An ultimate erosion operation is similar to the thinning operation, but instead of having thin connected scribbles we get small separated connected components. As a consequence of the reduction given by the ultimate erosion operator, only the essential information of the cells centrality is preserved, allowing us to have a concise representation of the center of the original object. A sample image and the results of its ultimate erosion can be observed in Figure 3.

III. FoxP3 cells Dataset

In this work, we propose a new dataset, composed by 10 annotated oral tissue sample images. The images contain a high resolution of 2048x1532 pixels, and were acquired by the Pontifícia Universidade Católica de Minas Gerais (PUC - Minas) dentistry laboratory. In each image, all of the FoxP3+ Treg cells are stained in a shade of brown. It is also possible to visualize cells in a non-brown color. However, cells of non-brown color won't be taken into account for this study for in future research the emphasis will be on pinpointing the ratio and evaluating the impact of the proportion of FoxP3+ cells in relation to non-brown cells to the presence of mouth cancer. We can observe an example of dataset image on Figure 1.

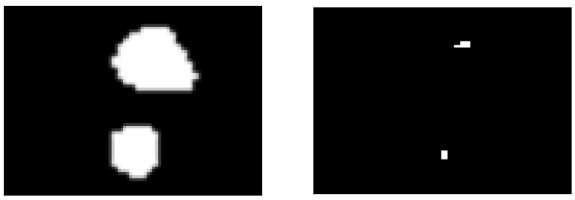

Fig. 3. **Object marker creation: Ultimate erosion onto binarized prediction weights.** The image on the left represents the activation map after binarization, where we applied a threshold to select as objects the set of pixels with the highest prediction weights (*grayscale value* $>$ *100*). The image on the right is the result after the ultimate erosion procedure, using the binarized image as input.

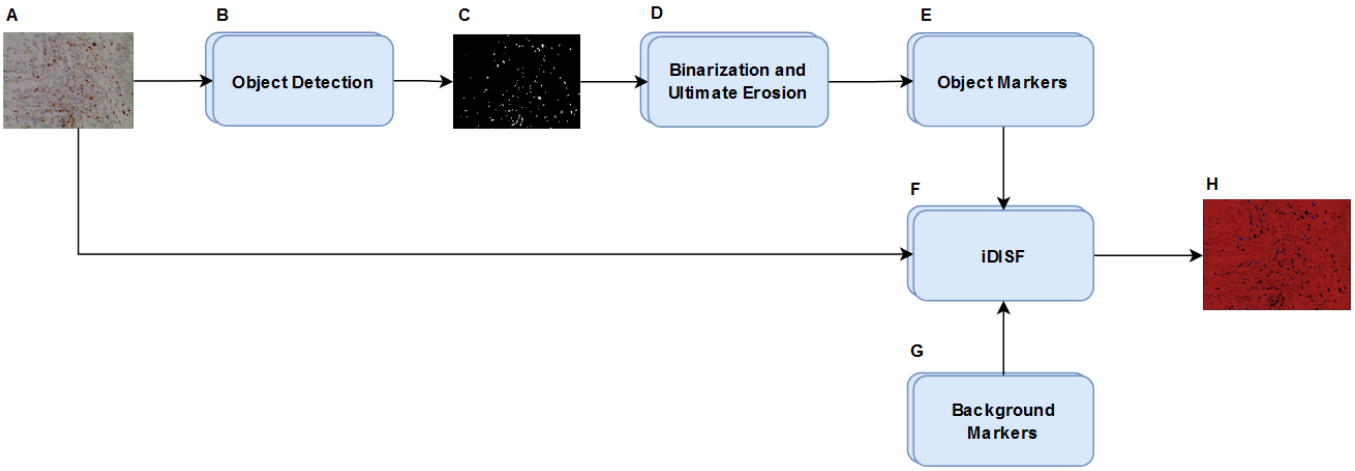

Fig. 4. **Complete Segmentation Pipeline. A** FoxP3+ cells dataset image, corresponding to the input of step **B**. **B** In the Object Detection step the input image is fed into LodeSTAR, whose output prediction weights are used next to generate the activation maps. **C** Binary activation map generated in step **B**, used as input in step **D**. **D** Ultimate Erosion procedure. In this step only the innermost pixels of the object is mantained, creating coordinates which will be used as object markers in iDISF (**E**). The object markers from step **E** will be inputed into iDISF to create seeds and, along with the background markers from step **G**, will be used to generate the segmentation in step **H**.

A. Samples and Imunohistochemistry

The procedure for acquiring this images is detailed in the following. To begin our sample extraction, paraffin-embedded biopsies from tissues previously collected for diagnostic purposes, with a clinical and histopathological diagnosis of oral leukoplakia, were included and evaluated by immunohistochemical staining. Streptavidin-biotin protocol was used for immunohistochemistry reaction. The serial sections of 3 μm in thickness of paraffin-embedded tissues were performed. The serial sections were deparaffinized, dehydrated, and antigen retrieval was carried out using Trilogy solution (Cell Marque, Rocklin, CA, USA) for 12 min at 98 °C. The samples were incubated in two baths of 0.3% hydrogen peroxidase for 15 min each for block out endogenous peroxidase activity. The specimens were incubated with monoclonal antibody FoxP3 (Abcam, clone:236A/E7, dilution 1:50), incubated at room temperature for 1 h. Detection was performed using the Reveal System (Spring bioscience, Pleasanton, CA, USA) incubated at room temperature for 30 min. The slides were subsequently exposed to 3,3-diaminobenzidine tetrahydrochloridechromogen (DAB, Sigma Chemical, St. Louis, USA, D5637). Mayer's hematoxylin was used for counterstaining. For reaction analysis, the slices were digitized with images captured by a digital camera attach to an Olympus BX51 optical microscope (Olympus Optical, Tokyo, Japan) interfaced to a computer, at a magnification of 400×. This study was approved by the Ethics Committee in Research of the Pontifícia Universidade Católica de Minas Gerais (PUC - Minas).

B. Ground-Truth

The dataset ground-truth consists of a collection of pre-segmented binary cell images. The iDISF platform was similarly employed during this procedure. However, cell marking was accomplished manually through its graphical interface under the supervision of dental specialists. The tool was chosen due to the high quality of its segmentation and the ease of later comparing the segmentation generated by our architecture. The ground truth and its original image can be seen in Figure 1. The dataset and the ground-truth will be publicly available.

IV. Methodology

Our pipeline is divided mainly in three steps, as illustrated in Figure 4: (i) **object detection**; (ii) **marker generation**; and (iii) **image segmentation**. Firstly we use the trained LodeSTAR deep learning model [11] to detect the location of cells in the input image (object detection). Then we apply morphological operations on its activation map in order to transformed it into a binary image containing the object markers (marker generation). Finally we use these object markers and the image into iDISF method to produce the cell segmentation (image segmentation). In the next subsections we explain in details the steps of this process.

A. Object Detection Architecture

In order to detect FoxP3+ cells positions, the LodeSTAR neural network proposed on [11] is used. This model consists on a single shot architecture created for the purpose of detecting microscopical objects. LodeSTAR achieves an accurate microscopic object detection by exploiting roto-translational equivariance. The architecture comprises a sequence of architectural components. Its first implementation level consists of three consecutive convolutional layers of size $3 \times 3 \times 32$, each employing ReLU activation. Subsequently, these convolutional layers are succeeded by a 2×2 max-pooling layer, followed by an arrangement of eight additional $3 \times 3 \times 32$ convolutional layers utilizing ReLU activation. The concluding element of this network is a singular $1 \times 1 \times 3$ convolutional layer with no activation function. Despite training with only one sample,

the network is able to generalize and detect multiple objects by removing the weighted global pooling layer and operating directly on the feature-maps, on the predicted object position map and on the weight map. The detection of the objects is obtained by acquiring the local maxima of a detection map, produced from the multiplication of the weight map and a measure of the local density of the object positions.

B. Network Training

As stated by the authors in [11], the model requires only a sample of the desired object so that the other cells are detected throughout the entire image. Therefore, our training set consists of a cropped image of size (50x50), obtained from one of the input images. As preprocessing, Gaussian shift, Gaussian blur and data normalization are applied sequentially before their insertion into the network, as shown in Figure 5. During the training stage, the network was trained several times, varying only the number of epochs by one in order to find the configuration that best detected. At the end of this process it was discovered that the variation of two to three epochs was enough to obtain a desirable detection. Due to the small size of the network, it was possible to carry out the training in just a few minutes. The whole process was tested in an Ubuntu environment with an octa-core CPU Intel Corei7-8550U.

C. Automatic Marker Creation

In iDISF, the marker generation process is completely manual, depending on the user to locate and mark the object and background seeds in order to produce the segmentation. Since user interaction is an expensive resource, in this work we propose an automatic seed extraction method to replace the user's marking in the interactive segmentation loop.

Let S'_O and S'_B be the novel set of object and background markers generated automatically, respectively. This approach is based on the use of LodeSTAR detection to generate the basis of the iDISF markers. We obtain its activation map and transform it into a binary image where the object detection is labelled with 255 and any other information is labelled with 0. To do so, we binarize the image obtained from the network weights with a threshold of 100. Before generating the markers we apply an ultimate erosion technique on the objects in order to reduce it into only few pixels.

This simplification process improves iDISF's performance by ensuring that only the necessary information to locate the position of the cells is used in order to create S'_O. S'_B is created in sequence using the borders of the image due to the rare amount of patterns similar to the FoxP3+ cells observed in these regions, mitigating possible marking errors. iDISF then expands the S'_B set by oversampling background seeds over the image using grid sampling. The goal is to retrieve the parts of the background that aren't part of S'_B and have different features (like color or texture) compared to any pixel in S'_B. We denote the expanded background seeds by S'^e_B.

Finally, iDISF processes the image and the seed inputs and outputs $Seg^I(S'_O, S'^e_B)$.

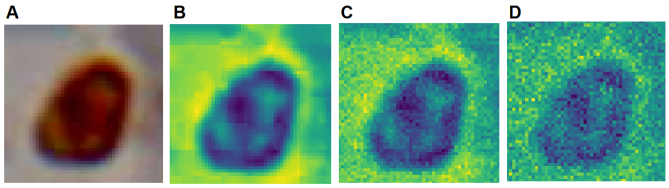

Fig. 5. **Training image preprocessing procedure. A** Original training image. **B** *Gaussian shift* is applied onto **A**. **C** *Gaussian blur* is applied onto **B**. **D** *Normalization* is applied onto **C**.

V. Experiments and Results

This section demonstrates that our proposed method can achieve satisfactory cell segmentation results across the entirety of the 10 images within our dataset. We evaluate the segmentation quality of our approach by employing three commonly used metrics in computer vision, the mean IoU (Intersection over Union) score and the Dice coefficient for assessing the segmentation quality, and the F1-score for assessing the acurracy rate of the cells detection module.

The IoU metric measures the degree of overlap between the segmented region generated by our approach and the region that should be segmented, referred as "ground truth". An IoU value of 0 denotes the absence of overlap between the generated segmentation and the ground truth, while an IoU value of 1 signifies an ideal scenario of perfect alignment between the two. Let A be the first image, B be the second image and N the amount of images, the mean IoU metric is calculated as show in Equation 1

$$MeanIoU = \frac{1}{N}\sum_{i=1}^{N}\frac{A_i \cap B_i}{A_i \cup B_i} \tag{1}$$

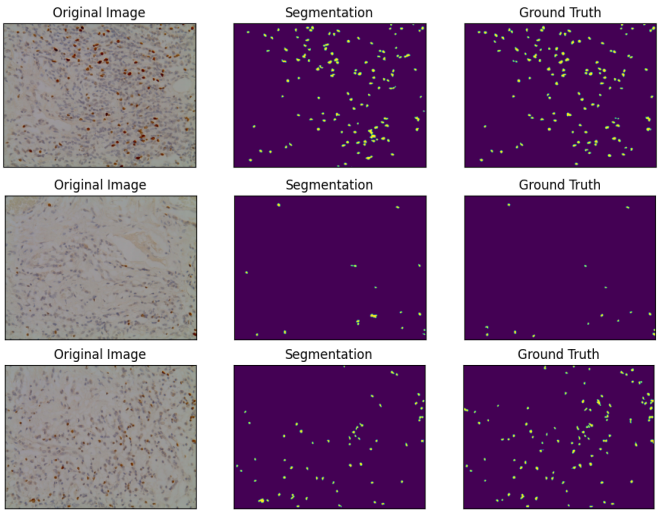

Fig. 6. **Comparison between the segmentation obtained through our method and the ground truth.** The first column shows examples of the FoxP3$^+$ cells dataset images. The second column exhibits the segmentation obtained through iDISF, after being fed the object markers generated by our approach. The third column presents the ground truth for said dataset images.

In turn, the Dice coefficient computes the intersection degree between two images. It is calculated as demonstrated in Equation 2:

$$MeanDice = \frac{1}{N}\sum_{i=1}^{N}\frac{2 \times |A_i \cap B_i|}{|A_i| + |B_i|} \tag{2}$$

Similarly to the IoU metric, a Dice coefficient that approaches 0 tends to have low similarity between the given images, while a coefficient index that approaches 1 tends to have a high similarity. Finally, the F1 score is computed as shown in Equation 3.

$$F1Score = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Recall} + \text{Recall}} \tag{3}$$

Let TP be the true positive values, FP be the false positive values and the FN be the false negative, the precision and recall are calculated as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

Our best configuration obtained a average IoU score of **0.52** and a mean Dice coefficient of **0.67** over the whole dataset. Regarding the cells detection phase, we achieved a a F1-Score of **0.73**. It was also computed that the object detection module had a **10%** false positive rate and a **30%** false negative rate. The amount of incorrectly detected cells have a direct impact on the segmentation quality scores, but it is important to note that we achieve these scores by training the object detection module with only one object instance, composed by a 50×50 image patch.

The Dice and IoU scores indicate a reasonable similarity between what was segmented and the expected ideal segmentation, specially taking into account the misdetected cells. In Figure 6 is possible to visualize the segmentation obtained in comparison with the ground truth in the form of a binary image. It is possible to perceive a considerable similarity between the images, which is maintained for all other images in the dataset. Despite the high similarity, the segmentation is still not perfect. For the correctly identified cells, we achieve very good pixel-level accuracy using iDISF, and we believe that with a better object detection, we will be able to achieve even better scores.

VI. Conclusion and future works

In this work we presented a novel method to segment FoxP3+ cells images in order to assist healthcare professionals in the task of identifying and counting potentially cancerous cells, requiring only a 50x50 training patch required to do so. We also propose a new dataset, composed by 10 mouth tissue images. Unfortunately, even if the results are quite satisfactory with the 10 images, the marker generation must be improved since the network is still not capable of perfectly identifying all the desired cells.

In future works, we plan to further study the creation of markers based on cell detection approaches, which can lead to an overall improvement on the final segmentation. Furthermore, we plan to study more robust methods to produce background markers, since we only use the frame of the image as an initial weak background marker.

Finally, we plan to collect more high-quality images in order to increase the FoxP3+ cells dataset. In special, we believe the proposed method can be used to help dental professionals annotate newly acquired images, thus speeding up the annotation process, which is usually very time-consuming.

Acknowledgment

The authors thank the Pontifícia Universidade Católica de Minas Gerais – PUC-Minas, the Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq (407242/2021-0, 306573/2022-9) and Fundação de Apoio à Pesquisa do Estado de Minas Gerais – FAPEMIG (APQ-01079-23).

References

[1] S. Sakaguchi, M. Miyara, C. M. Costantino, and et al, "Foxp3+ regulatory t cells in the human immune system," *Nature Reviews Immunology*, vol. 10, pp. 490–500, 2010.

[2] Szylberg, D. Karbownik, and A. Marszałek, "The role of foxp3 in human cancers," *Anticancer Research*, vol. 36, no. 8, pp. 3789–3794, 2016.

[3] O. Stasikowska-Kanicka, M. Wagrowska-Danilewicz, and M. Danilewicz, "Immunohistochemical analysis of foxp3+, cd4+, cd8+ cell infiltrates and pd-l1 in oral squamous cell carcinoma," *Pathology Oncology Research*, vol. 24, no. 3, pp. 497–505, July 2018.

[4] N. Hiraoka, K.-I. Onozato, T. Kosuge, and et al, "Prevalence of foxp3+ regulatory t cells increases during the progression of pancreatic ductal adenocarcinoma and its premalignant lesions," *Clinical Cancer Research*, vol. 12, pp. 5423–5434, 2006.

[5] N. Kobayashi, N. Hiraoka, W. Yamagami, and et al, "Foxp3+ regulatory t cells affect the development and progression of hepatocarcinogenesis," *Clinical Cancer Research*, vol. 13, pp. 902–911, 2007.

[6] N. Altini, A. Brunetti, E. Puro, M. G. Taccogna, C. Saponaro, F. A. Zito, S. De Summa, and V. Bevilacqua, "Ndg-cam: Nuclei detection in histopathology images with semantic segmentation networks and grad-cam," *Bioengineering*, vol. 9, no. 9, 2022. [Online]. Available: https://www.mdpi.com/2306-5354/9/9/475

[7] M. S. E. Rabbi, N. Ironside, J. A. Ozolek, R. Singh, L. Pantanowitz, and G. K. Rohde, "Transport-based morphometry of nuclear structures of digital pathology images in cancers," 2023.

[8] I. B. Barcelos, F. Belém, P. Miranda, A. X. Falcão, Z. K. do Patrocínio, and S. J. F. Guimarães, "Towards interactive image segmentation by dynamic and iterative spanning forest," in *International Conference on Discrete Geometry and Mathematical Morphology*, Springer. Springer International Publishing, 2021, pp. 351–364.

[9] 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. 2274–2282, 2012.

[10] A. X. Falcao, J. Stolfi, and R. de Alencar Lotufo, "The image foresting transform: Theory, algorithms, and applications," *IEEE transactions on pattern analysis and machine intelligence*, vol. 26, no. 1, pp. 19–29, 2004.

[11] B. Midtvedt, J. Pineda, F. Skärberg, E. Olsén, H. Bachimanchi, E. Wesén, E. K. Esbjörner, E. Selander, F. Höök, D. Midtvedt, and G. Volpe, "Single-shot self-supervised object detection in microscopy," *Nature Communications*, vol. 13, no. 1, p. 7492, 2022. [Online]. Available: https://doi.org/10.1038/s41467-022-35004-y