Bounding Box is all you need: Learning to Segment Cells in 2D Microscopic Images via Box Annotations

Nabeel Khalid[1,2][0000−0001−9274−3757], Maria Caroprese[5][0009−0009−2170−1459], Gillian Lovell[4][0009−0004−5180−9704], Daniel A Porto[6][0000−0002−1021−2467], Johan Trygg[3,7][0000−0002−4239−6520], Andreas Dengel[1,2][0000−0002−6100−8255], and Sheraz Ahmed[1][0000−0002−4239−6520]

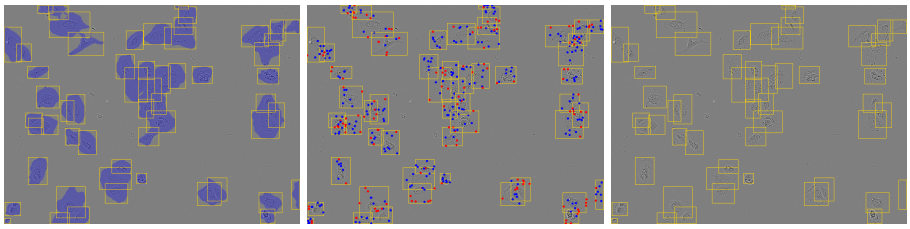
[1] German Research Center for Artificial Intelligence (DFKI) GmbH, Kaiserslautern 67663, Germany
firstname.lastname@dfki.de
[2] RPTU Kaiserslautern–Landau, Kaiserslautern 67663, Germany
[3] Sartorius Corporate Research, Sweden
[4] Sartorius, BioAnalytics, Royston, United Kingdom
[5] Sartorius, Digital Solutions, Royston, United Kingdom
[6] Sartorius, BioAnalytics, Ann Arbor, United States
firstname.lastname@sartorius.com
[7] Computational Life Science Cluster (CLiC), Umeå University, Sweden

Abstract. Microscopic imaging plays a pivotal role in various fields of science and medicine, offering invaluable insights into the intricate world of cellular biology. At the heart of this endeavor lies the need for accurate identification and characterization of individual cells within these images. Deep learning-based cell segmentation, which involves delineating cells from complex microscopic images, is pivotal for cell analysis. It serves as the foundation for extracting meaningful information about cell morphology, spatial organization, and interactions. However, traditional deep-learning models for cell segmentation require extensive and expensive annotation masks for each cell in the image, posing a significant challenge. To address this issue, this study introduces CellBoxify, a novel pipeline that streamlines cell instance segmentation. Unlike traditional methods, CellBoxify operates solely on bounding box annotations, making it approximately seven times faster than manual segmentation mask annotation for each cell. The proposed approach's effectiveness is evident in its performance on the LIVECell dataset, a well-known resource for cell segmentation research. Achieving 83.40% of the fully supervised performance on this dataset demonstrates the efficacy of the proposed method.

Keywords: cell segmentation · weakly supervised · medical imaging · deep learning · bounding box annotations.

1 Introduction

In microscopic analysis, cells are crucial for understanding biology and diseases, guiding advancements in medicine. Cell segmentation is crucial for disease diagnosis, research, and drug discovery, aiding in identifying abnormal phenotypes and studying dynamics. However, deep learning methods [3,8–10,17,18] typically demand extensive fully annotated data, making the process time-consuming and costly, leading to a need for weakly supervised approaches.
The LIVECell dataset [3] is pivotal in cell biology research, boasting over 1.6 million cells with an unparalleled density averaging 313 cells per image, a significant leap from datasets like EVICAN [17]. Annotation of cells in microscopic images presents distinct challenges due to their small scale, complexity, variability, and noise, unlike natural images. Dense cell cultures like BV2 and morphologically intricate types such as SH-SY5Y further complicate manual annotation, impacting accurate annotation. Additionally, the task is compounded by the dataset's high cell volume, with an average annotation time of 46 seconds per cell in the LIVECell dataset.
In cell biology, the absence of annotated data limits the application of supervised deep-learning models for precise cell segmentation. Annotating cellular data demands substantial resources and expertise, leading to a scarcity of labeled datasets. Consequently, much available cellular data remains unannotated, restricting opportunities to enhance segmentation algorithms. To overcome these challenges, this study introduces CellBoxify, a deep learning-based pipeline for cell instance segmentation utilizing only bounding boxes for training. Figure 1 illustrates the disparity in annotation requirements between the fully supervised method (Figure 1a), the other weakly supervised methods [7, 11] (Figure 1b), and the proposed weakly supervised approach Figure (1c), CellBoxify. Previous



(a) Full Supervised (b) Weak Supervised [7, 11] (c) CellBoxify

Fig. 1: **Full supervised (a) vs. Weak supervised (b) vs. CellBoxify (c)** input training samples. The fully supervised method requires a full mask, whereas the weakly supervised approaches, Point2Mask [11] and PACE [7], need bounding box and point annotations. The blue and red points represent whether the point lies on the cell or outside, respectively. The proposed approach, CellBoxify, requires only the bounding box for training.

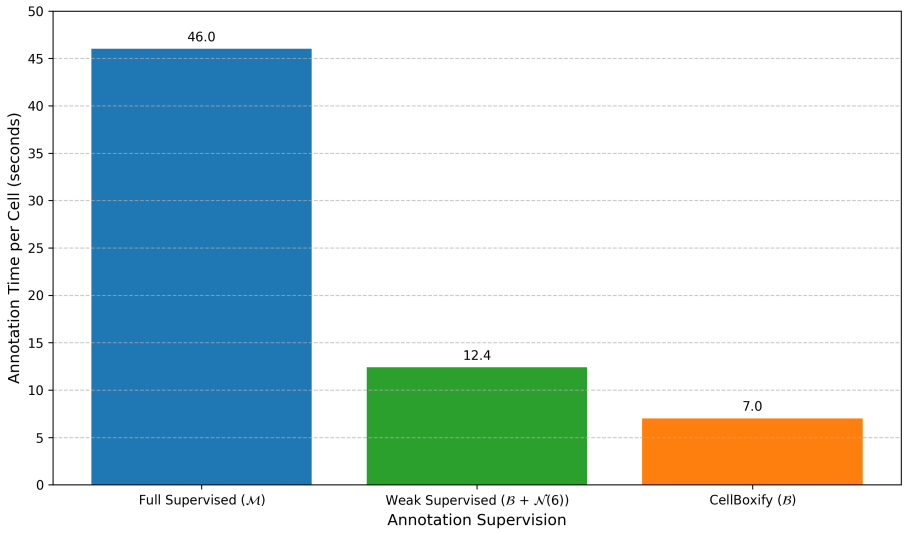


Fig. 2: Annotation times per cell for three approaches: Full Supervised ($\mathcal{M}$), Weak Supervised ($\mathcal{B}$+$\mathcal{N}$(6)), and CellBoxify ($\mathcal{B}$). $\mathcal{M}$ is for fully supervised Mask supervision, $\mathcal{N}$ represents Point supervision, and $\mathcal{B}$ represents Box supervision. CellBoxify, using only bounding box annotation, reduces annotation time by over 6.5 times compared to fully supervised.

weakly supervised approaches for cell segmentation, such as Point2Mask [11] and PACE [7], involve a two-step annotation process. Initially, annotators draw bounding boxes for each cell, followed by the automated generation of points within these boxes. These points are then automatically assigned labels based on the available segmentation mask for each cell, determining their classification as foreground (cell) or background. However, this approach faces challenges in distinguishing points accurately in scenarios with densely clustered cells, leading to potential labeling errors and increased annotation time. While the annotation times per cell for different approaches are compared in Figure 2, it's essential to note that the automated labeling process may not accurately replicate real-world scenarios where manual decisions are required. Particularly in practical situations with densely clustered cells, accurately annotating points can be challenging and may result in increased annotation time. In contrast, the proposed CellBoxify approach streamlines the annotation process by requiring only single-stage annotations, specifically bounding box annotations, which take approximately 7 seconds per cell. By eliminating the need to annotate individual points within bounding boxes, CellBoxify simplifies the workflow, reduces annotation overhead, and offers a more efficient solution for weakly supervised cell segmentation. The main contributions of this study are as follows:

1. This study introduces CellBoxify, a deep learning pipeline for cell instance segmentation using only bounding box supervision in microscopic images leveraging Mask R-CNN [5], Feature pyramid Network with ResNet-50 [6], along with point-based unary and distance-based pairwise losses to optimize instance segmentation polygon predictions [19].
2. Evaluation of the proposed approach using the LIVECell dataset. Achieved 83.40% of the fully supervised performance using CellBoxify with a significant reduction in the time required for data annotation.
3. Further evaluation conducted on a per cell culture basis to identify challenges associated with specific morphological characteristics and their impact on segmentation performance.

2 Related Work

2.1 Fully Supervised Cell Segmentation

In the past decade, there has been remarkable progress in fully supervised deep learning-based cell analysis, particularly utilizing segmentation masks for each cell. A significant milestone occurred with the introduction of the U-net architecture by Ronneberger et al. in 2015 [16]. Despite being trained on only 35 images, the U-net model outperformed all competitors in the 2015 ISBI cell tracking and segmentation challenge. This breakthrough not only demonstrated the efficacy of deep learning in cell segmentation but also catalyzed a series of significant advancements in image-based cellular research. Subsequent to U-net, pioneering algorithms such as CellPose [18], DeepCeNS [9], DeepCIS [10], and DeepMuCS [8] emerged, further pushing the boundaries of fully supervised cell segmentation techniques.

2.2 Weakly Supervised Cell Segmentation

Acquiring the annotations required for training fully supervised deep learning models poses a significant challenge due to its laborious and intricate nature. In response, researchers have proposed weakly supervised or semi-supervised learning approaches to alleviate the annotation burden. Weakly supervised techniques, such as image tags [21] and point annotations [7,11] have been explored in the realm of cell segmentation. Notably, Khalid et al. introduced Point2Mask [11] and PACE [7] as methods for weakly supervised cell segmentation, which utilize multiple points along with a bounding box for each cell. These approaches achieved remarkable performance, attaining 99.2% and 99.8% accuracy compared to fully supervised methods, using 6- and 8-point labels and bounding boxes, respectively. However, these methods entail a two-step annotation process for cell segmentation. Initially, annotators delineate bounding boxes for each cell, followed by the automatic generation of points within these boxes. Subsequently, points are automatically labeled based on the available segmentation mask, determining whether they belong to the foreground (cell) or background.

One notable challenge associated with this approach is the difficulty in distinguishing points in scenarios where cells are densely clustered together. This can lead to errors in labeling and potentially increase the annotation time. Furthermore, it's essential to acknowledge that the annotation time reported for these approaches may not accurately reflect real-world scenarios. The automatic labeling of points as belonging to the foreground or background relies on the segmentation mask available for each cell. In practical settings, where the segmentation mask is not readily available, experts must manually determine whether a point lies inside or outside the cell. This manual decision-making process can be particularly challenging, especially when cells are densely clustered together, potentially resulting in increased annotation time and inaccuracies.

3 CellBoxify: The Proposed Approach

Figure 3 illustrates the system overview of the proposed pipeline for bounding box-based cell instance segmentation, CellBoxify. The proposed method is composed of Mask R-CNN [5], Feature Pyramid Network (FPN) with ResNet-50 [6], along with point-based unary and distance-based pairwise losses to optimize instance segmentation polygon predictions [19]. The proposed pipeline is composed of three main modules: Backbone Network, Region Proposal Network, and Prediction Head.

3.1 Backbone Network

The purpose of this block is to extract feature maps from the input image at different scales. The feature extraction module of the proposed methodology is composed of Feature Pyramid Network [12] along with ResNet-50 [6]. FPN employs a pyramid scheme to extract features from images, utilizing deep convolutional networks (CNNs) for this purpose. It combines lower resolution, but semantically strong features with higher resolution, yet semantically weak features. This is achieved through a multi-step process involving a bottom-up pathway, a top-down pathway, and lateral connections. In the bottom-up pathway, a standard feed-forward CNN architecture is employed to compute a hierarchy of features, generating feature maps at various scales. These feature maps serve as the basis for subsequent operations. The top-down pathway utilizes the output of each convolutional layer from the ResNet-50 network, integrating them via lateral connections to construct higher-resolution layers from the semantically rich layers. To address aliasing effects resulting from upsampling, a 3x3 convolution operation is applied to each merged map as part of the final stage of FPN. This operation helps refine the feature maps, ensuring that the final output accurately captures the essential characteristics of the input image.

3.2 Region Proposal Network

After extracting multi-scale features from the backbone network, the next step involves passing these features through a Regional Proposal Network (RPN),

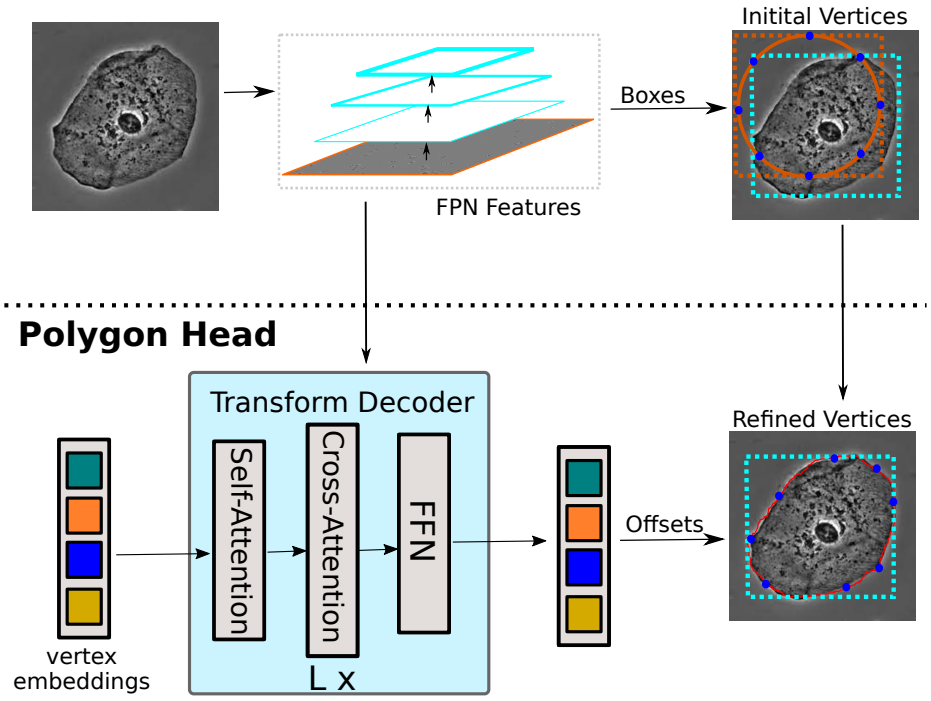


Fig. 3: **System overview of the CellBoxify.** Multiscale features are extracted from the input image by a backbone network. A box predictor is attached to these features to obtain bounding boxes. The polygon head predicts the polygon for each box, which is trained with box annotation only.

as proposed by Ren et al. [15]. The primary objective of the RPN is to identify regions within the image that potentially contain objects and align them with ground truth annotations. This process begins by generating anchor boxes across the input image, which are then matched to ground truth annotations using the Intersection over Union (IoU) metric. Anchors with an IoU greater than a predefined threshold (typically 0.7) are linked to ground truth boxes and classified as foreground objects. Those with an IoU between 0.3 and 0.7 are considered background, while those with an IoU below 0.3 are ignored. However, the default anchor strides and aspect ratio parameters designed for detecting and segmenting objects in datasets like MS-COCO [13] often overlook small cell instances present in datasets like LIVECell [3]. Extensive experimentation led to the selection of anchor sizes and aspect ratios tailored to this specific task. Unlike MS-COCO and other standard image datasets, the LIVECell dataset features exceedingly small cells, particularly in BV-2 cell cultures. The optimal anchor parameters, detailed in Section 5, were carefully chosen to suit this task. After assigning anchor boxes that match the shapes of ground truth boxes, the next step involves calculating anchor deltas, which denote the distance between

ground truth and anchors. In the final stage of the RPN, 3,000 region proposal boxes are selected from the predicted boxes using non-maximum suppression [1].

3.3 Prediction Head

Following the feature extraction and region proposal stages, the prediction head in the CellBoxify pipeline is tasked with generating accurate polygonal masks for each detected object. Unlike traditional methods that rely on handcrafted energy functions, the proposed approach leverages deep learning techniques to accomplish polygon-based instance segmentation with only bounding box supervision.

Point-based Unary Loss To ensure that all vertices of the predicted polygon fall within the ground-truth bounding box, a point-based unary loss is introduced. This loss function minimizes the discrepancy between the predicted bounding box and the ground truth using the complete intersection over union metric:

$$L_u = 1 - CIoU(b_c, b)$$

where CIoU$(\cdot, \cdot)$ represents the complete intersection over union [20].

Distance-aware Pairwise Loss While the point-based unary loss ensures tight bounding box alignment, it may fail to accurately fit object boundaries. Therefore, a distance-aware pairwise loss is proposed, consisting of both local and global terms.

Local Pairwise Term Object boundaries often exhibit local color variation in images [4]. To enforce local consistency, a local pairwise loss based on windows is introduced. By reformulating the polygon prediction as a classification problem, the predicted polygon is encouraged to align with image edges within a local region.

$$E = \sum_{(p,q)\in\hat{\Omega}_k(i,j)} w[(i,j),(p,q)] \cdot |U'_C(i,j) - U'_C(p,q)|$$

where $\hat{\Omega}_k(i,j)$ denotes the adjacent pixels within a $k \times k$ window at position (i,j), and $w[(i,j),(p,q)]$ measures the affinity of two pixels by color distance.

Global Pairwise Term To mitigate the influence of local noise, a global pairwise loss is introduced. This loss encourages homogeneous color regions within and outside the predicted polygon, resulting in smoother and more accurate segmentation boundaries [2].

$$L_{gp} = \sum_{(x,y)\in\Omega} \|I(x,y) - u_{in}\|^2 \cdot U'_C(x,y) + \sum_{(x,y)\in\Omega} \|I(x,y) - u_{out}\|^2 \cdot (1 - U'_C(x,y))$$

where u_{in} and u_{out} indicate the average image color inside and outside the predicted polygon, respectively.

Clipping Strategy To manage memory constraints during training, a clipping strategy is employed. This strategy involves resizing the predicted polygon to a fixed size using bilinear interpolation and using RoIAlign [5] to crop and resize the image based on the coordinates of the ground-truth box. By reducing memory requirements, this strategy enhances the practicality of CellBoxify for users with limited computational resources.

Joint Loss Function Finally, the proposed approach integrates the point-based unary loss (L_u) and the distance-aware pairwise loss (L_{lp} and L_{gp}) into a joint loss function:

$$L_{polygon} = \alpha L_u + \beta L_{lp} + \gamma L_{gp}$$

where α, β, and γ are the modulated weights for each loss term. During training, this loss function guides the network to predict accurate object polygons with only bounding box supervision. The joint loss function enables the network to effectively learn from bounding box annotations and generate precise polygonal masks, facilitating high-quality instance segmentation without the need for pixel-level supervision.

4 Dataset

In cell biology research, the LIVECell dataset, outlined by Edlund et al. (2021) [3], is an invaluable resource. This dataset boasts 1.6 million cells across 5,239 meticulously chosen images, featuring eight distinct cell cultures and an impressive average of 313 cells per image. Notably, its high cell density exceeds that of comparable datasets like EVICAN (Schwendy et al., 2020) [17]. Despite its complexity, LIVECell offers researchers a unique opportunity to explore densely populated cellular environments, reflecting real-world conditions. For this study, only bounding box annotations are used for training, streamlining the annotation process, and maximizing the dataset's utility for comprehensive cellular analysis.

5 Experimental Setup

Two distinct experimental setups were devised to assess the efficacy of the CellBoxify pipeline for weakly supervised cell segmentation. The first setup, termed "Comparative Analysis of Supervision Methods: LIVECell," compares CellBoxify, Point2Mask, and a Fully Supervised method using various annotation strategies on the LIVECell dataset. The second setup, titled "Comparative Analysis of Supervision Methods: Per Cell Culture," evaluates how different training supervisions impact performance across individual cell cultures. Models trained in the first setup are then evaluated on test sets corresponding to specific cell cultures to assess their performance.

Compared to natural scene images, microscopic images often suffer from low contrast, which presents challenges in accurately distinguishing cells from the

background. In low-contrast scenarios, cell boundaries become unclear, posing challenges for edge identification and local pairwise loss computation critical for the optimization of the proposed approach. As a result, predicted polygons struggle to capture precise cell boundaries, hindering effective vertex coordinate optimization. To mitigate this, a preprocessing step using Contrast Limited Adaptive Histogram Equalization (CLAHE) [14] is applied to enhance image contrast in the LIVECell dataset during CellBoxify training. CLAHE redistributes pixel intensities to improve visual clarity while preserving overall image appearance, preventing over-enhancement common in traditional histogram equalization methods.

All training scenarios, including CellBoxify, Point2Mask, and Fully Supervised approaches, utilized Mask R-CNN [5] with ResNet-50 [6] as the underlying framework. Training for CellBoxify employed the ADAM solver with a base learning rate of 0.0001 and a momentum of 0.9. Anchor sizes and aspect ratios were carefully configured based on the pixel area of cells in the images, with sizes set to 8, 16, 32, 64, 128, and aspect ratios to 0.5, 1, 2, 3, 4 across all experimental settings to ensure adaptability to diverse cell sizes and shapes. Data augmentation included random horizontal flipping to mitigate overfitting, and multi-scale data augmentation, where image sizes were randomly adjusted to promote robustness and generalization. Specifically, image sizes were resized to one of the following lengths: 440, 480, 520, 580, or 620 pixels, ensuring exposure to various scales during training.

The performance of the proposed and other training supervision methods was evaluated following the standard COCO evaluation protocol [13], with adjustments specified in [3] regarding the maximum number of detections and area ranges. Evaluation checkpoints were selected based on higher validation average precision.

5.1 Experimental Setting 1: Comparative Analysis of Supervision Methods: LIVECell

In Experimental Setting 1, the performance of three different supervision methods: Full Mask, Point2Mask, and the proposed weakly supervised method CellBoxify are evaluated. Each method is trained on the LIVECell training data using distinct annotation strategies: Full Mask involves training with segmentation masks for each cell, Point2Mask is trained using both bounding boxes and points (6 in this case), while CellBoxify relies solely on bounding box supervision.

Table 1 presents a comprehensive overview of the performance of each method on the complete LIVECell test set. The Full Mask ($\mathcal{N}$) and Point2Mask ($\mathcal{B}+\mathcal{N}(6)$) trained models achieve segmentation AP scores of 43.90% and 43.53%, respectively. In contrast, CellBoxify ($\mathcal{B}$) achieves a segmentation AP score of 36.61%, which corresponds to 83.40% of the performance of the Full Mask trained model.

Table 1: Segmentation Performance Comparison on the LIVECell Test Set: Full Mask vs. Point2Mask vs. CellBoxify

Train	**AP**		**AP50**		**AP75**		**APs**		**APm**		**APl**	
Supervision	**Det.**	**Seg.**	**Det.**	**Seg.**	**Det.**	**Seg.**	**Det.**	**Seg.**	**Det.**	**Seg.**	**Det.**	**Seg.**
Full Mask ($\mathcal{M}$)	43.12	43.90	78.94	78.07	43.26	45.75	44.31	42.30	43.01	43.33	47.01	51.92
Point2Mask ($\mathcal{B} + \mathcal{N}(\mathbf{6})$)	43.32	43.53	79.69	78.18	43.31	44.93	44.54	42.06	43.31	43.31	46.97	51.52
CellBoxify ($\mathcal{B}$)	40.16	36.61	77.12	72.59	38.92	34.93	41.49	36.67	39.29	33.93	41.90	41.90

5.2 Experimental Setting 2: Comparative Analysis of Supervision Methods: Per Cell Culture

In this experimental setting, the primary objective is to explore the correlation between the morphological characteristics of different cell cultures and their segmentation performance under various supervisions. The aim is to identify the specific attributes of individual cell cultures that contribute to segmentation challenges or facilitate easier segmentation across the three different training strategies.

Figure 4 showcases the segmentation outcomes obtained for each cell culture, presenting a comparative analysis of the performance achieved by the Full Mask, Point2Mask, and CellBoxify approaches. Among the results, CellBoxify demonstrates its best performance with the SkBr3 (59.90%) and BV-2 (47.74%) cell cultures, closely matching the performance of models trained with Full Mask and Point2Mask methods. Conversely, CellBoxify exhibits lower performance with the SH-SY5Y cell culture, displaying a segmentation performance gap of 11.5% and 11.3% compared to the Full Mask and Point2Mask methods, respectively.

6 Analysis and Discussion

This section provides a comprehensive analysis of the findings obtained from the two experimental setups, shedding light on their broader implications. In the first experimental setting, "Comparative Analysis of Supervision Methods: LIVECell," the results of employing three different annotation supervision techniques on the complete LIVECell test set are presented. The results from Table 1 demonstrate that, despite being trained solely with bounding box supervision rather than using full masks or a combination of bounding boxes and points, the proposed approach, CellBoxify, achieves an impressive 83.40% of the performance obtained by models trained with full masks. This signifies the efficacy of CellBoxify in achieving competitive segmentation results while significantly reducing the annotation overhead.

The outcomes of the second experimental setting, namely "Comparative Analysis of Supervision Methods: Per Cell Culture," explore the correlation between

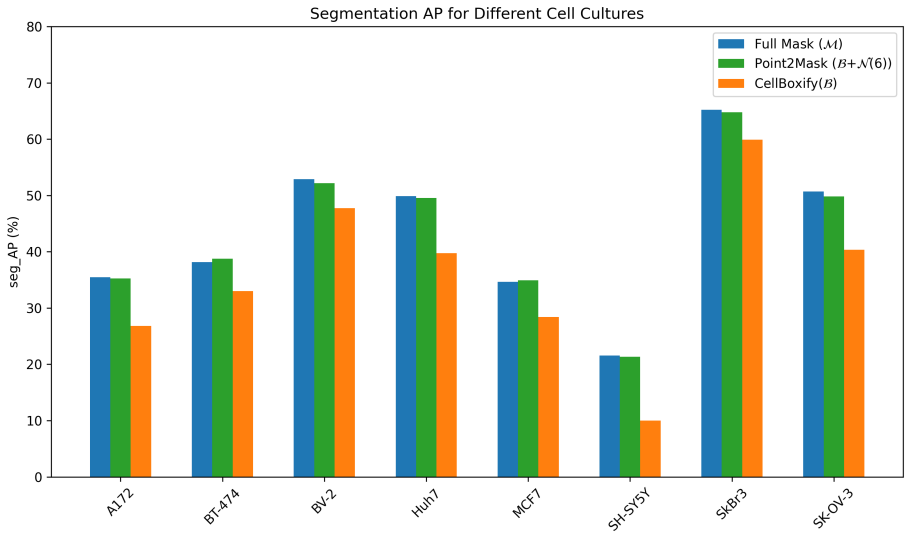


Fig. 4: Segmentation Performance Comparison Across Cell Cultures: Full Mask vs. Point2Mask vs. CellBoxify

the morphological characteristics of different cell cultures and their segmentation performance under various supervision methods. Specifically, the results from this setting provide insights into the areas for improvement in the current pipeline for weakly supervised cell segmentation. From Figure 4, it is evident that the best performance is observed for cell cultures with cells in small area ranges, such as BV-2 and SkBr3. Conversely, the worst performance is observed across the SH-SY5Y cell culture, characterized by its neuronal cells with unique morphologies compared to other cell types. Neuronal cells often exhibit highly asymmetric and concave shapes due to their branching neurites, posing challenges for conventional cell segmentation models [18]. The neuronal structure of cells results in lower contrast, making it more difficult to delineate cell boundaries, especially when using weakly supervised approaches for training. These findings suggest a potential direction for future research, where targeted preprocessing techniques could be applied to enhance the performance of cell cultures with complex morphologies, such as those containing neuronal cells.

Figure 5 illustrates the inference results on various samples using models trained on Full Mask (purple column), Point2Mask (blue column), and the proposed approach CellBoxify (green column). The solid yellow lines represent the ground truth masks for each cell, while the dotted red lines depict the model predictions. Each row demonstrates the qualitative performance of different training supervisions on identical images from distinct cell cultures for comparative analysis. The segmentation average precision score at the IoU threshold of 0.5 (AP50) is indicated above every prediction sub-image. The best-performing method for

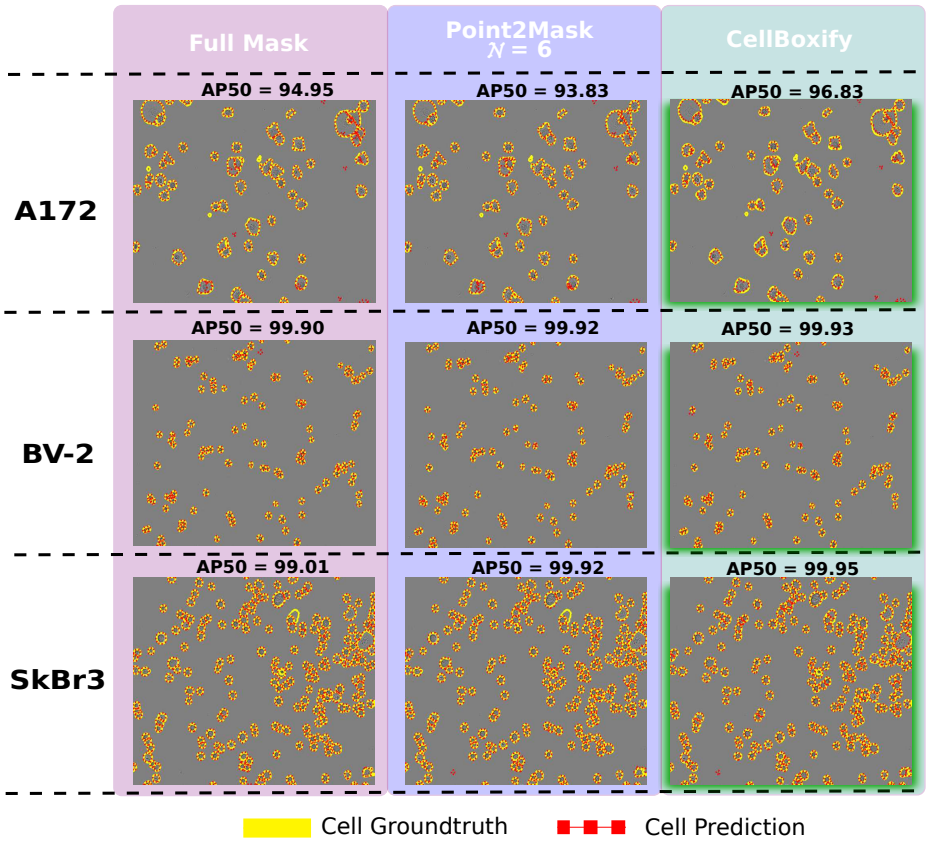


Fig. 5: Comparison of Adequate Segmentation Results Using Different Training Methods: Full Mask (purple), Point2Mask (blue), and CellBoxify (green) illustrate ground truth masks (solid yellow) and model predictions (dotted red), with AP50 scores above each prediction sub-image. The best-performing method is highlighted with a green glow.

each image is highlighted with a green glow around the resulting image. For the A172 image, CellBoxify achieves the highest AP50 score of 96.83%, compared to 94.95% and 93.83% for the Full Mask and Point2Mask methods, respectively. Similar performance can be observed for the BV-2 and SkBr3 images, where the proposed approach achieves AP50 scores of 99.93% and 99.95% for segmentation. Figure 6 showcases inadequate inference results on various samples using the three different training supervisions. For the BT-474 image, the AP50 performance gap between CellBoxify and the Full Mask is 6.79%, with the Full Mask performing better. For the SH-SY5Y image, the performance gap increases to 23.74%, indicating that CellBoxify struggles to perform well on the SH-SY5Y cell culture. However, for the SkBr3 image, CellBoxify outperforms both the Full Mask and Point2Mask, achieving an AP50 score of 51.93%.

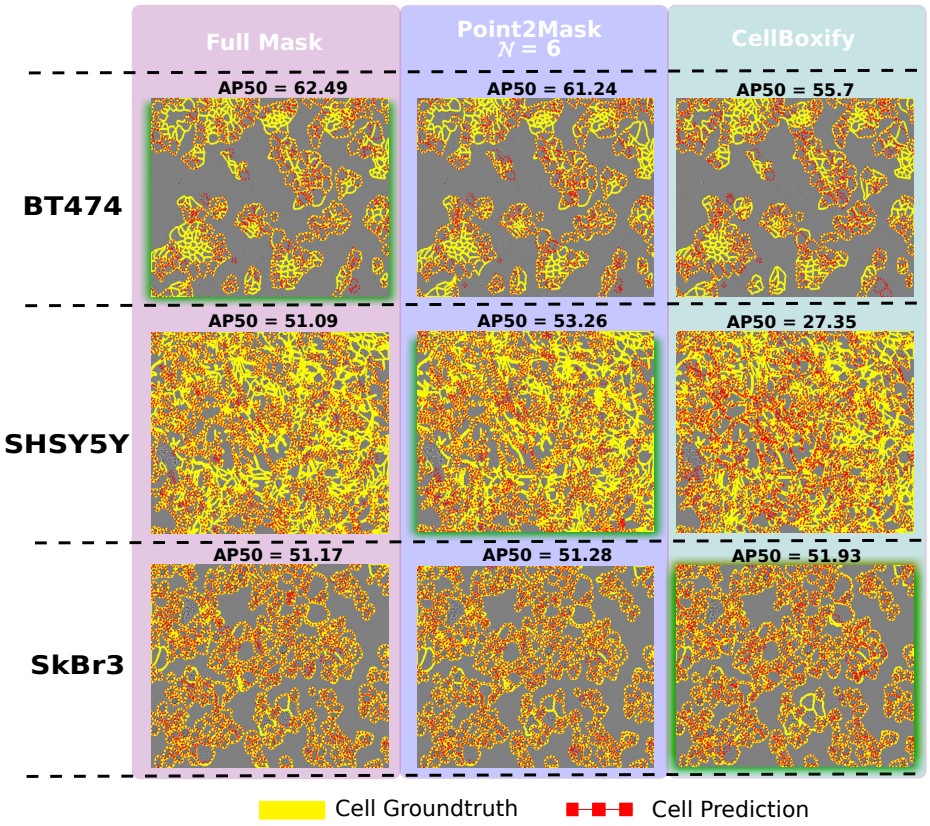


Fig. 6: Comparison of Inadequate Segmentation Results Using Different Training Methods: Full Mask (purple), Point2Mask (blue), and CellBoxify (green) illustrate ground truth masks (solid yellow) and model predictions (dotted red), with AP50 scores above each prediction sub-image. The best-performing method is highlighted with a green glow.

The introduction of CellBoxify represents a breakthrough in microscopic image analysis. The other weakly supervised approaches like Point2Mask [11] and PACE [7] adopt a two-step annotation process, necessitating both bounding box annotations and point annotations. However, in scenarios where cells are densely clustered together, this approach can become intricate, potentially leading to labeling errors and increased annotation time. By employing bounding box-based segmentation, CellBoxify streamlines the annotation process, saving up to 85% of annotation time while maintaining commendable performance. This paradigm shift accelerates research progress and facilitates deeper exploration of complex cellular processes. CellBoxify offers a practical solution for large-scale image analysis tasks, balancing segmentation accuracy with annotation efficiency. Beyond its time-saving advantages, CellBoxify's versatility lies in its ability to accu-

rately segment cells using only bounding box annotations. This feature makes it ideal for scenarios where obtaining precise segmentation masks is challenging or resource-intensive.

7 Conclusion

CellBoxify introduces a groundbreaking approach to cell segmentation, utilizing only bounding box annotations for network training. Achieving 83.40% of the Full Mask training supervision's performance, CellBoxify demonstrates its efficacy while saving approximately 85% of the time typically spent on annotating masks. This innovative method not only streamlines the annotation process but also enhances efficiency and scalability in biomedical image analysis. The findings of this study have significant implications for biologists and medical professionals, offering a potential time-saving solution in data labeling processes that could greatly expedite advancements in medicine and disease diagnosis. By showcasing the efficacy of the proposed weakly supervised approach, this research not only addresses the challenge of annotation time and costs but also alleviates the expertise burden on biologists required for manual cell boundary delineation. With a wealth of unlabeled image-based cellular data available, the application of the proposed pipeline for semi-automated annotation holds promise for efficiently annotating large datasets for further analysis and research in cell segmentation. Looking forward, there is potential for further advancements by refining preprocessing techniques tailored to specific cell types or imaging conditions. Such efforts could optimize segmentation performance and unlock new possibilities for comprehensive cell analysis in biomedical research and beyond.

References

1. Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2018)
2. Chan, T.F., Vese, L.A.: Active contours without edges. IEEE Transactions on image processing (2001)
3. Edlund, C., Jackson, T.R., Khalid, N., Bevan, N., Dale, T., Dengel, A., Ahmed, S., Trygg, J., Sjögren, R.: Livecell—a large-scale dataset for label-free live cell segmentation. Nature methods (2021)
4. Gonzalez, R., Woods, R.: Digital image processing, addison-wesle y longman publishing co. Inc, Boston, MA, USA (2001)
5. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision (2017)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2016)
7. Khalid, N., Froes, T.C., Caroprese, M., Lovell, G., Trygg, J., Dengel, A., Ahmed, S.: Pace: Point annotation-based cell segmentation for efficient microscopic image analysis. In: International Conference on Artificial Neural Networks. Springer (2023)

8. Khalid, N., Koochali, M., Rajashekar, V., Munir, M., Edlund, C., Jackson, T.R., Trygg, J., Sjögren, R., Dengel, A., Ahmed, S.: Deepmucs: A framework for co-culture microscopic image analysis: From generation to segmentation. In: 2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI). IEEE (2022)
9. Khalid, N., Munir, M., Edlund, C., Jackson, T.R., Trygg, J., Sjögren, R., Dengel, A., Ahmed, S.: Deepcens: An end-to-end pipeline for cell and nucleus segmentation in microscopic images. In: 2021 International Joint Conference on Neural Networks (IJCNN). IEEE (2021)
10. Khalid, N., Munir, M., Edlund, C., Jackson, T.R., Trygg, J., Sjögren, R., Dengel, A., Ahmed, S.: Deepcis: An end-to-end pipeline for cell-type aware instance segmentation in microscopic images. In: 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI). IEEE (2021)
11. Khalid, N., Schmeisser, F., Koochali, M., Munir, M., Edlund, C., Jackson, T.R., Trygg, J., Sjögren, R., Dengel, A., Ahmed, S.: Point2mask: A weakly supervised approach for cell segmentation using point annotation. In: Medical Image Understanding and Analysis: 26th Annual Conference, MIUA 2022, Cambridge, UK, July 27–29, 2022, Proceedings. Springer (2022)
12. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2017)
13. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. Springer (2014)
14. Pizer, S.M., Amburn, E.P., Austin, J.D., Cromartie, R., Geselowitz, A., Greer, T., ter Haar Romeny, B., Zimmerman, J.B., Zuiderveld, K.: Adaptive histogram equalization and its variations. Computer vision, graphics, and image processing (1987)
15. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. arXiv preprint arXiv:1506.01497 (2015)
16. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. Springer (2015)
17. Schwendy, M., Unger, R.E., Parekh, S.H.: Evican–a balanced dataset for algorithm development in cell and nucleus segmentation. Bioinformatics (2020)
18. Stringer, C., Wang, T., Michaelos, M., Pachitariu, M.: Cellpose: a generalist algorithm for cellular segmentation. Nature Methods (2020)
19. Yang, R., Song, L., Ge, Y., Li, X.: Boxsnake: Polygonal instance segmentation with box supervision. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)
20. Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., Ren, D.: Distance-iou loss: Faster and better learning for bounding box regression. In: Proceedings of the AAAI conference on artificial intelligence (2020)
21. Zhou, Y., Zhu, Y., Ye, Q., Qiu, Q., Jiao, J.: Weakly supervised instance segmentation using class peak response. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)