Luca Rettenberger*, Friedrich Rieken Münke, Roman Bruch, and Markus Reischl

Mask R-CNN Outperforms U-Net in Instance Segmentation for Overlapping Cells

https://doi.org/10.1515/cdbme-2023-1084

Abstract: U-Net is the go-to approach for biomedical segmentation applications. However, it is not designed to segment overlapping objects, a challenge Mask R-CNN has shown to have great potential in. Yet, Mask R-CNN receives little attention in biomedicine. Hence, we evaluate both approaches on a publicly available biomedical dataset. We find that Mask R-CNN outperforms U-Net in segmenting overlapping cells and achieves comparable performance if they do not intersect. Our study provides valuable decision support to practitioners in selecting an appropriate method when solving instance segmentation tasks using deep learning, as well as important insights into enhancing the accuracy of such approaches in biomedical image analysis.

Keywords: Instance Segmentation, Machine Learning, Computer Vision, Deep Learning

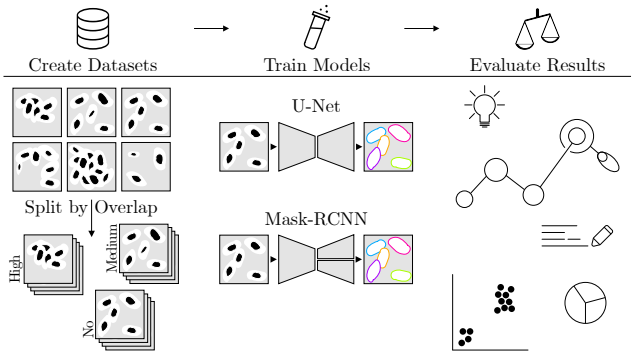

Fig. 1: Overview. We subdivide a biomedical dataset based on instance overlap and train U-Net and Mask R-CNN on each split. After that, we assess the neural networks' ability to learn from overlapping instances.

1 Introduction

Instance segmentation is a task that involves delineating individual objects within an image. It is an essential task in biomedical research, having crucial roles in all areas from disease diagnosis over drug discovery to cell behavior analysis. In recent years, Deep Learning (DL) has emerged as a powerful tool for image analysis, promising to tackle the most challenging segmentation tasks. But despite the success and potential of DL, many challenges are still considered very difficult to solve. In biology, segmenting overlapping cells is such a demanding and recurring challenge [6].

One of the most popular DL architectures for cell segmentation is U-Net [9], which has established itself as the go-to approach in biomedical applications. It is known for its simplicity and efficiency [7, 8, 11]. However, despite its success, U-Net still faces challenges, such as learning instance segmentations, which requires significant modifications to the original architecture [10]. Alternative approaches exist but have received little attention in biomedical applications to date. One is Mask R-CNN [3], which combines object detection and instance segmentation to produce precise masks around each object in an image. It has shown great potential when segmenting cells [2]. However, there are only a limited number of studies comparing the performance of Mask R-CNN and U-Net in semantic segmentation [1, 12], and none that focus on instance segmentation. In this work, we compare Mask R-CNN with U-Net for cell segmentation. Both approaches are evaluated on a publicly available biomedical dataset containing cells obtained from cervical cytology [6]. A visual summary of our work is given in Fig. 1.

Our study highlights the importance of exploring alternative approaches to address the challenges of biomedical image analysis and provides decision support to help practitioners decide when to use which method. The analysis incorporates both qualitative evaluations and quantitative metrics, including a novel measure that examines the intensity of overlapping instances in a dataset, which can be utilized to further enhance and simplify the decision process. Moreover, the paper includes a ready-to-use implementation of the discussed methods, making it easy to replicate the experiments and build upon the findings. Overall, the paper's contributions offer important insights for improving the accuracy and efficiency of instance segmentation methods in biomedical image analysis. Our experiment pipeline, the used methods, and metrics are open-source and available at: https://github.com/lrettenberger/maskrcnn-vs-unet-for-instance-segmentation.

***Corresponding author: Luca Rettenberger,** Institute for Automation and Applied Informatics, Karlsruhe Institute of Technology, Eggenstein-Leopoldshafen, Germany, e-mail: luca.rettenberger@kit.edu
Friedrich Rieken Münke, Roman Bruch, Markus Reischl, Institute for Automation and Applied Informatics, Karlsruhe Institute of Technology, Eggenstein-Leopoldshafen, Germany

2 Method

The main idea of our work is to critically examine the established way of segmenting cells in the biomedical field that usually employs U-Nets, particularly in cases where instances overlap. To accomplish this, we conduct an experiment that compares Mask R-CNN [3] with a modified version of the U-Net architecture for instance segmentation of cells [10]. To obtain expressive results, we employ a biomedical dataset that we divide into a number of sub-datasets based on the number of overlapping instances. The results are evaluated, both graphically as well as by using quantitative metrics, including a novel measure designed to evaluate performance in relationship with overlap ratio (for an overview see Fig. 1).

2.1 Dataset

The dataset consists of 945 synthetic cervical cytology images, which were generated using cells extracted from 16 non-overlapping field-of-view images obtained from four specimens [6]. The images vary in number of cells present and degree of overlap and are designed to replicate the characteristics of real cervical cytology images. The masks for each cell in an image are provided as instance segmentation masks. Fig. 2 displays two samples of the dataset. All samples in the dataset have the dimension 512×512 pixels and are grayscale images.

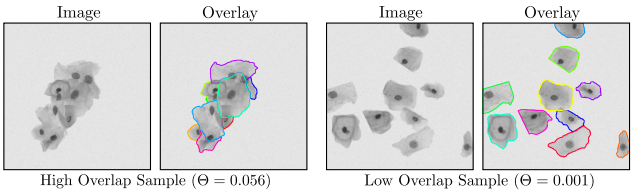

Fig. 2: Two samples of the dataset used in this work including the ground truth masks marked in color. The first sample contains many overlapping cells and the second one almost none.

2.2 Experiment Design

Our conceptual approach uses a dataset $\mathcal{D}$ of length M with different levels of overlapping objects. To accurately assess the impact of overlap, $\mathcal{D}$ is partitioned into n sub-datasets $\mathcal{S} = \{\mathcal{S}_1, \mathcal{S}_2, \dots \mathcal{S}_n\}$ of equal size, each containing different amounts of instance overlap within the samples. Every ground truth instance segmentation mask $\mathcal{D}_i,\ i < M$ is given as an array of binary masks with C instances, height H, and width W. So the whole dataset has the form $\text{shape}(\mathcal{D}) = [M, C, H, W]$.

Tab. 1: The sub-datasets used in this work. Θ is the amount of overlap within the cells and #Train / #Test is the number of train and test samples.

Sub-Dataset	Θ	#Train	#Test
$\mathcal{S}_1$	0%	131	24
$\mathcal{S}_2$	1%	131	30
$\mathcal{S}_3$	12%	131	28
$\mathcal{S}_4$	19%	131	28
$\mathcal{S}_5$	26%	131	27
$\mathcal{S}_6$	37%	125	28

Let $\mathcal{D}^f$ be the class frequency describing for every pixel coordinate h, w, how many classes overlap at this position

$$\mathcal{D}^f_{i,h,w} = \sum_{c=1}^{C} \mathcal{D}_{i,c,h,w}. \tag{1}$$

Using $\mathcal{D}^f$, we calculate the percentage of pixels that belong to multiple classes (overlaps) relative to the overall area that all masks take for a segmentation mask $\mathcal{D}_i$ as

$$\Gamma_i = \frac{\sum_{h=1}^{H}\sum_{w=1}^{W}[\mathcal{D}^f_{i,h,w} > 1]}{\sum_{h=1}^{H}\sum_{w=1}^{W}[\mathcal{D}^f_{i,h,w} > 0]}, \tag{2}$$

where $[\bullet]$ denotes the Iverson bracket [4]. With the overlap criterion Γ_i we create the n sub-datasets $\mathcal{S}$ by sorting $\mathcal{D}$ based on the overlapping area and creating n equal-sized bins by calculating the size of each bin as $\lceil |\mathcal{D}|/(n+1) \rceil$, where $|\mathcal{D}|$ denotes the number of samples in $\mathcal{D}$ (see Fig. 1).

To evaluate the neural networks based on the intensity of overlap in a sample, we introduce the novel overlapping measure Θ that describes how much overlap is present in an image. The maximum amount of overlap occurs if all instances are completely aligned on top of each other, while the minimum amount happens if there is no overlap at all. The maximum and minimum possible overlap frequencies $\mathcal{D}^{f\uparrow}$ and $\mathcal{D}^{f\downarrow}$ for a sample $\mathcal{D}_i$ are determined as

$$\mathcal{D}^{f\uparrow} = \sum_{h=1}^{H}\sum_{w=1}^{W}[\mathcal{D}^f_{i,h,w} > 0] * C, \tag{3}$$

$$\mathcal{D}^{f\downarrow} = \sum_{h=1}^{H}\sum_{w=1}^{W}[\mathcal{D}^f_{i,h,w} > 0]. \tag{4}$$

With $\mathcal{D}^{f\uparrow}$ and $\mathcal{D}^{f\downarrow}$ we calculate Θ as

$$\Theta = \frac{(\sum_{h=1}^{H}\sum_{w=1}^{W}\mathcal{D}^f_{i,h,w}) - \mathcal{D}^{f\downarrow}}{\mathcal{D}^{f\uparrow} - \mathcal{D}^{f\downarrow}}, \tag{5}$$

with Θ in the range $(0.0 \leq \Theta \leq 1.0)$. High Θ values indicate high overlap. Tab. 1 shows the resulting splits $\mathcal{S}$ when dividing the dataset (Sec. 2.1) into equal-sized bins after sorting by Θ.

Tab. 2: The results for each sub-dataset. The metric listed is the Aggregated Jaccard Index (AJI+) [5]. δ is the percentage increase from U-Net to Mask R-CNN, expressed as a percentage of the U-Net value (relative gap).

Sub-Dataset	U-Net	Mask R-CNN	δ
$\mathcal{S}_1$	95% $\pm$ 0.5	90% $\pm$ 1.0	-5%
$\mathcal{S}_2$	84% $\pm$ 3.3	87% $\pm$ 0.7	$+3\%$
$\mathcal{S}_3$	78% $\pm$ 1.5	81% $\pm$ 0.8	$+4\%$
$\mathcal{S}_4$	64% $\pm$ 1.9	72% $\pm$ 0.5	$+13\%$
$\mathcal{S}_5$	54% $\pm$ 2.3	63% $\pm$ 0.9	$+17\%$
$\mathcal{S}_6$	41% $\pm$ 4.3	55% $\pm$ 1.4	$+34\%$

3 Results

3.1 Architecture, Training, and Implementation

We use the Dice Loss as the objective function and the Adam optimizer with a learning rate of 0.002 (U-Net) or 0.0001 (Mask R-CNN) in all experiments, which are values found to be optimal in a preceding hyperparameter search. We do not employ data augmentation for training. All sub-datasets are randomly divided into 80% / 20% splits of training and validation data. The samples are normalized to be in the range $[0, 1]$. For the U-Net-based approach, we predict Euclidean distance maps with subsequent seed-based watershed post-processing for segmenting instances [10]. Our implementation of Mask R-CNN follows the original introduction [3] and is configured with two classes: cell and background. Early stopping and learning rate scheduling are employed. Both, U-Net and Mask R-CNN are implemented in PyTorch Lightning. We conduct the training using an NVIDIA GeForce RTX 3090 GPU and an AMD Ryzen 9 5950X 16-Core 3.40GHz CPU. To avoid initialization effects and ensure reliable metrics, we repeat all experiments four times with random seeds and report the mean results along with their corresponding standard deviation. For evaluation, we use a test split of the corresponding sub-dataset.

3.2 Experiments

Tab. 2 shows the results of the experiment, split by sub-datasets. Our findings reveal that U-Net outperforms Mask R-CNN if the cells do not intersect. However, if there is a slight overlap between the objects, the performance of U-Net decreases considerably. With more challenging sub-datasets Mask R-CNN also generates less fitting masks, but the performance drops less rapidly. With the sub-dataset containing the largest overlap, this even results in a relative gap of 34% between U-Net and Mask R-CNN. These results are consistent

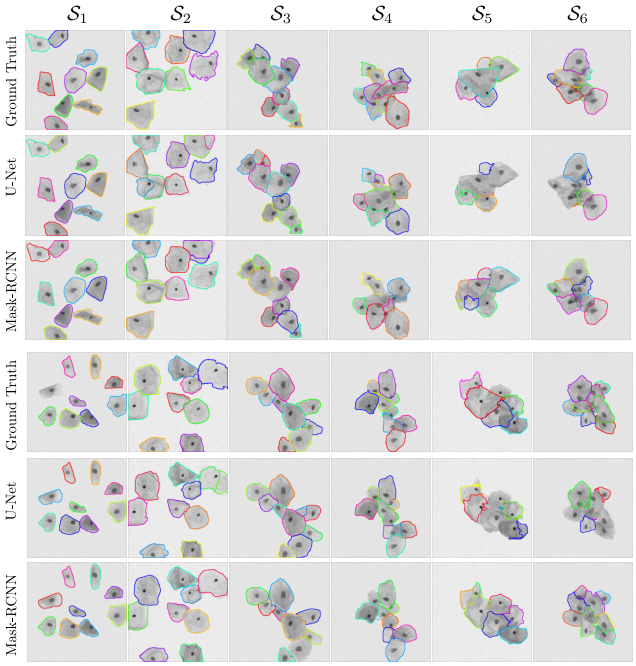

Fig. 3: Two samples from the test data from each sub-dataset with the ground truth instance segmentation masks, U-Net predictions, and Mask R-CNN predictions.

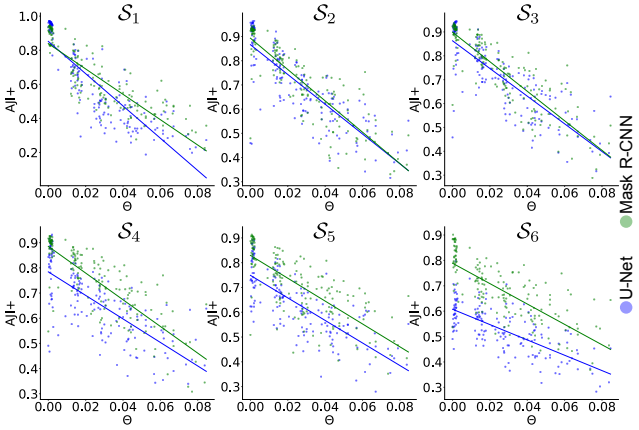

Fig. 4: Scatter plot with Θ on the x-axis and the AJI+ [5] metric on the y-axis. The best U-Net and Mask R-CNN model for training with each sub-dataset is shown. Each point is one sample from the combined test data, so all samples from all dataset splits. For both methods, a linear regression line is plotted over the scatter points.

with what can be seen when looking at the generated instance segmentation masks. Fig. 3 shows two samples of each sub-dataset with the ground truth segmentation masks, as well as the generated ones by U-Net and Mask R-CNN. For $\mathcal{S}_1$ both approaches generated precise masks. With $\mathcal{S}_2$ the masks still appear to be fitting, but U-Net already displays some slight inaccuracies. Looking at $\mathcal{S}_3$ similar behavior is apparent. Even for $\mathcal{S}_4$, the dissimilarities are not yet immediately noticeable.

But with $\mathcal{S}_5$, the strengths of Mask R-CNN become noticeable. Here, U-Net often does not recognize cells or cannot find the boundaries correctly. This intensifies even more with $\mathcal{S}_6$.

The scatter plots in Fig. 4 evaluate the best models for each sub-dataset. Unlike the previous evaluations, all test sets are merged to evaluate every model across the entire range of Θ. The results show that Mask R-CNN improves if trained on samples with higher Θ. Interestingly, even when trained on $\mathcal{S}_1$ (no overlap), Mask R-CNN achieves noticeably better performance than U-Net on samples with high Θ. For $\mathcal{S}_2$ and $\mathcal{S}_3$ (slight overlap) Mask R-CNN and U-Net were about equal in performance. However, for $\mathcal{S}_4$, $\mathcal{S}_5$, and $\mathcal{S}_5$, Mask R-CNN consistently outperforms U-Net. Interestingly, if the networks are trained on data with high Θ, Mask R-CNN maintains good performance with samples drawn from the entire spectrum of Θ which is not the case with U-Net.

4 Discussion

In this study, we observed noticeable differences between U-Net and Mask R-CNN in their performance in instance segmentation tasks. Specifically, we found that U-Net excelled in scenarios where the instances were distinct and non-overlapping. However, if the dataset contains overlapping instances, U-Net's performance decreases rapidly, making Mask R-CNN the more suitable solution. With increasing levels of overlap, Mask R-CNN remains considerably more stable and provides much more precise segmentation masks. Interestingly, if trained on challenging masks with much overlap, Mask R-CNN maintains good performance on simple samples as well. This suggests that Mask R-CNN is capable of learning spatial correlations from challenging samples that can aid in identifying objects in more trivial ones. In contrast, U-Net is not capable of this. We suspect this is a structural problem as the U-Net is limited to making one prediction per pixel, which leads to contradictions if several objects are in the same position. This seems to restrict U-Net in its capabilities so considerably that it's not capable of segmenting non-overlapping cells if trained solely on high-overlap samples.

5 Conclusion

We recognize that U-Net, while widely used in biomedical applications, is not well suited for instance segmentation if the objects are overlapping. Our study compares U-Net with the alternative approach of Mask R-CNN on a synthetic cervical cytology images dataset with many overlapping cells. We observe that while U-Net had a slight advantage when there is no overlap, this advantage diminishes with even slight intersections of objects, and Mask R-CNN's advantage grows with increasing overlap. Overall, our findings suggest that Mask R-CNN is a promising alternative to the widely used U-Net architecture, particularly in scenarios with many intersections of objects. Future research could compare Mask R-CNN and U-Net on further datasets with different overlap levels and assess other instance segmentation methods to enhance future biomedical applications.

Acknowledgement

This work was supported by the HoreKa Supercomputer through the Ministry of Science, Research, and Art, Baden-Württemberg and by the Helmholtz Association Initiative and Networking Fund on the HAICORE@KIT partition.

References

[1] Alfaro E, et al. A brief analysis of u-net and mask r-cnn for skin lesion segmentation. In: Proceedings of the IEEE International Work Conference on Bioinspired Intelligence (IWOBI). 2019, 000123–000126.

[2] Fujita S, et al. Cell detection and segmentation in microscopy images with improved mask r-cnn. In: Proceedings of the Asian Conference on Computer Vision. 2020, 58–70.

[3] He K, et al. Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision. 2017, 2961–2969.

[4] Iverson KE. A programming language. In: Proceedings of the Spring Joint Computer Conference. 1962, 345–351.

[5] Kumar N, et al. A dataset and a technique for generalized nuclear segmentation for computational pathology. *IEEE Transactions on Medical Imaging* 2017;36:1550–1560.

[6] Lu Z, et al. Evaluation of three algorithms for the segmentation of overlapping cervical cells. *IEEE Journal of Biomedical and Health Informatics* 2016;21:441–450.

[7] Rettenberger L, et al. Annotation efforts in image segmentation can be reduced by neural network bootstrapping. *Current Directions in Biomedical Engineering* 2022;8:329–332.

[8] Rettenberger L, et al. Self-supervised learning for annotation efficient biomedical image segmentation. *IEEE Transactions on Biomedical Engineering* 2023;.

[9] Ronneberger O, et al. U-net: Convolutional networks for biomedical image segmentation. In: 18th International Conference on Medical Image Computing and Computer-assisted Intervention (MICCAI). 2015, 234–241.

[10] Scherr T, et al. Cell segmentation and tracking using cnn-based distance predictions and a graph-based matching strategy. *PLOS ONE* 2020;15:e0243219.

[11] Schutera M, et al. Methods for the frugal labeler: Multi-class semantic segmentation on heterogeneous labels. *PLOS ONE* 2022;17:e0263656.

[12] Zhao T, et al. Comparing u-net convolutional network with mask r-cnn in the performances of pomegranate tree canopy segmentation. In: Proceedings of the Society of Photo-Optical Instrumentation Engineers (SPIE), volume 10780. 2018, 210–218.