2022 International Conference on Artificial Intelligence in Information and Communication (ICAIIC) | 978-1-6654-5818-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICAIIC54071.2022.9722616

A Comparison of YOLO and Mask-RCNN for Detecting Cells from Microfluidic Images

1st Mehran Ghafari
Dept. of Computer Science
& Engineering
U. of Tennessee at Chattanooga
Chattanooga, TN, U.S.A.
ryg668@mocs.utc.edu

2nd Daniel Mailman
Dept. of Computer Science
& Engineering
U. of Tennessee at Chattanooga
Chattanooga, TN, U.S.A.
daniel-mailman@utc.edu

3rd Parisa Hatami
Dept. of Computer Science
& Engineering
U. of Tennessee at Chattanooga
Chattanooga, TN, U.S.A.
qxy699@mocs.utc.edu

4th Trevor Peyton
Dept. of Computer Science
& Engineering
U. of Tennessee at Chattanooga
Chattanooga, TN, U.S.A.
qtx464@mocs.utc.edu

5th Li Yang
Dept. of Computer Science
& Engineering
U. of Tennessee at Chattanooga
Chattanooga, TN, U.S.A.
li-yang@utc.edu

6th Weiwei Dang
Dept. Molecular & Human Genetics
Huffington Ctr. on Aging
Baylor Coll. of Medicine
Houston, U.S.A.
weiwei.dang@bcm.edu

7th Hong Qin
SimCenter, Dept. of Computer Science & Engineering
U. of Tennessee at Chattanooga
Chattanooga,TN, U.S.A.
hong-qin@utc.edu

***Abstract*—As an effective model to study aging, the budding yeast *Saccharomyces cerevisiae* has revealed aging mechanisms that are shared with human aging. Yeast cell lifespan can be measured in replicative lifespans (RLS) - the number of cell divisions from a single mother cell before dying. However, counting yeast cell divisions from microscopic images is a tedious task. Here, we address this challenge with computer vision object detection. We compared two deep learning methods, YOLO and MASK R-CNN to detect cells from microfluidic images. We concluded that YOLO is more sensitive at detecting cells, whereas MASK-RCNN is more informative on cell sizes. Therefore, both methods are useful for automatic microfluidic image analysis.**

***Index Terms*—machine learning, instance segmentation, cell detection, cellular aging.**

I. Introduction

Computer Vision (CV) approaches in recent years have led to advancements in many fields including medical, civil, surveillance, auto, etc [1]. There is a tremendous demand for CV in healthcare as many diagnoses and disease treatments rely on medical imaging [2]. Objects' appearances in images are associated with many features, most notably volume, dimensionality, color, resolution, and moving object demeanor. [3], [4].

This study analyses the effectiveness of CV techniques for microfluidic cell detection (MFCD). In MFCD images, cells are: visually extremely similar, extremely close together (often sharing boundaries), and often overlap due to the image being a map of 3 dimensions to 2 dimensions. These factors make cell detection a challenging task. Other factors which contribute to the difficulty of MFCD are uneven illumination, low contrast, low resolution, out-of-focus images, and varying foreground/background intensities [5]. The core task of object detection in general and MFCD in particular is segmentation - distinguishing object borders - into local and global regions [6], [7]. MFCD images contain hundreds of cells to distinguish using CV segmentation methods. Precision is required, especially in the identification of overlaps [8]. Segmentation methods and models rely on image pixel characteristics as well as sub-sectioning. Various methods and approaches have been implemented to improve segmentation efficiency.

Many segmentation models are based on a convolutional neural networks (CNNs). Based on preliminary literature examination, we chose two CNN-based models - You Only Look Once (YOLO) [9] and Mask R-CNN [10] - to evaluate for the task of MFCD.

YOLO uses a single CNN, predicts multiple bounding boxes, and determines class probability for each available image bounding box. YOLO is very fast and does not need a complex pipeline, since it relies on regression analysis. The model potentially runs at 45 frames per second (FPS) without batch processing requirements - meaning it is also capable of processing stream video in near-real-time. The model uses a simple down-sampling method which has the advantage of learning complex depth features of images using residual blocks. [11] used YOLO-based system to achieve 99.7% accuracy detecting mass located in the breast. YOLO's main drawback is using bounding boxes (rather than extracting

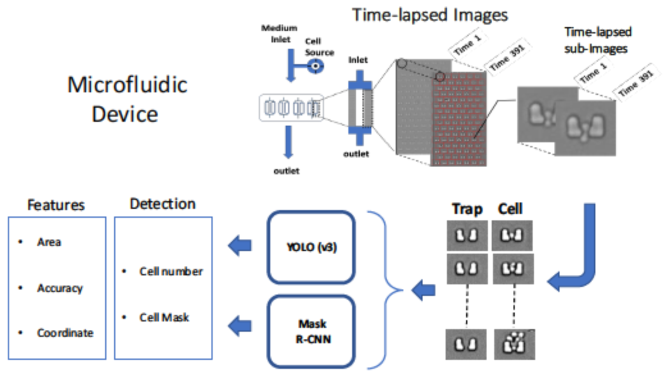

Fig. 1. Microfluidic device and image pre-processing steps. YOLO and Mask R-CNN are applied to partitioned microfluidic images. Each model's output yields object detection and feature extraction.

shape/contour details) in approximating target positions.

Mask R-CNN is an instance-segmentation method. It is a regional CNN model that generates detected object masks, increasing accuracy in contour detection and determining shape information [12]. However, some problems were reported that Mask R-CNN has a Resnet-101 [13] as its backbone which makes it a deep neural network. Hence, it requires more computational space for each training dataset. To address this problem, modified Mask R-CNN (Resnet-86) uses fewer backbone layers for vehicle and pedestrian detection [14]. Furthermore, [15] demonstrates that Mask R-CNN has poor performance for segmentation in comparison to U-net.

Details of the comparison of YOLO to Mask R-CNN follow in the remainder of this study. Section II addresses methods. Section III compares and discusses the two methods. Section IV summarizes the research.

II. Method

We used an Ubuntu 18.04.4 with Intel Xeon processor with 10 cores, 64GB of RAM, and nVidia RTX 2080 Ti GPU.

A. Dataset

The dataset is experimental results obtained from microfluidic HYAA chips [16]. Grayscale images were acquired by a microscope (Olympus IX-81) equipped with a camera (Olympus DP72 CCD). The temperature was set at 86°F. 391 time-lapse microfluidic images were taken at 10-minute intervals over a 96-hour period. On average, each image contains 104 silicon-made traps with rows of 6 or 7 traps. (Fig.1). Since the direct-object detection methods performed poorly on cell detection due to microfluidic low image resolution (grayscale 1280x960), we cropped traps by partitioning images into sub-images based on the number of available traps on each image. This approach is an effective technique for improving accuracy as well as generating more datasets without data augmentation.

B. Annotation process

We used 2 datasets. The first dataset (used for cell detection) contained 100 training sub-images and 30 test sub-images. The second dataset (used for feature extraction) contained 100 training sub-images and 40664 test sub-images. Sub-images for the first dataset were randomly selected from a batch containing a maximum of 5 cells per image.

We used "Microsoft VoTT Tool" and "Image-J" for image annotation. Mask R-CNN annotation is polygon-based. YOLO (bounding box) annotation format is [x,y,w,h], where (x,y) is the bounding box centroid, w is the width, and h is the height. Training-set sub-image dimension is 60X60. We used 60X60 and 512X512 sub-image size for the cell detection test dataset. The larger images were made using cubic interpolation.

C. YOLO

YOLO takes an image and estimates a confidence level for each detected object. YOLO' strategy is to reframe object detection as a single regression problem from image pixels to bounding box and classification probabilities. Fig 2a shows YOLO network architecture where the input image is 60x60, scaled up to 448x448x1. The next section is the DarkNet Architecture which is a CNN based on GoogleNet architecture [17]. DarkNet transforms image dimensions from 448x448x1 to 7x7x1024. Further, 2 full-connected neural networks are applied to the model with 2 outputs (2b): object bounding box (including object score) and class probability. In the entire YOLO network, the down-sampling of the network is based on setting the convolution stride hyperparameter to 2 without applying the pooling layer. The loss function consists of classification loss for the class probability and localization loss for the confidence level and bounding box which are both based on the squared error (sum).

[ht]

The improved version of this model is YOLOv2 [18], YOLOv3 [19], and YOLOv4 [20]. This work mainly focuses on YOLOv3, and all results are based on version 3 of this model.

D. Mask R-CNN

Region-Based CNN (R-CNN) is used for semantic segmentation and object detection and builds on other CNN models. The baseline models - Fast R-CNN [21], Faster R-CNN [22], and Fully Connected Network (FCN) [23] - are robust, pliable, fast-training, and conceptually intuitive. Mask R-CNN is based on Faster R-CNN. Mask R-CNN outperforms traditional semantic segmentation models by offering instance segmentation, including object mask. Fig 3 illustrates the varieties of R-CNN architecture. The salient differences among the models is summarized as follows.

In Fig 3a, multiple region features (size, shape, texture, color) are determined via multiple deep CNNs (e.g., AlexNet [24]) and fed separately to the bounding box offset regressor and the support vector machine (SVM) object classifier. In Fig 3b, the CNN region output is consolidated with a Region-Of-Interest (ROI) pooling layer. The consolidated data is fed to the regressor and the classifier enabling association of class labels to ROIs. In Fig 3c, multiple region proposals are eliminated in favor of using the CNN output as input to a Region Proposal

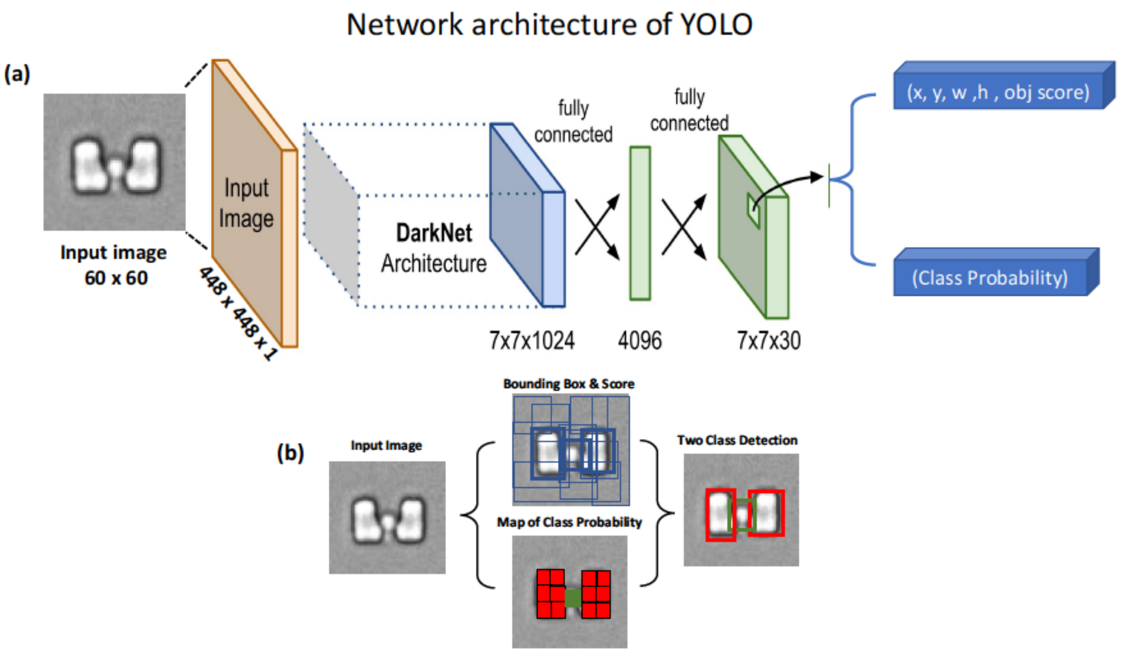

Fig. 2. YOLO architecture. (a) YOLO architecture with 60x60 image dimensions which scaled up to 448x448x1. The output contains bounding box information, object score, and object class. (b) Minimizing bounding box error with the map of class probability.

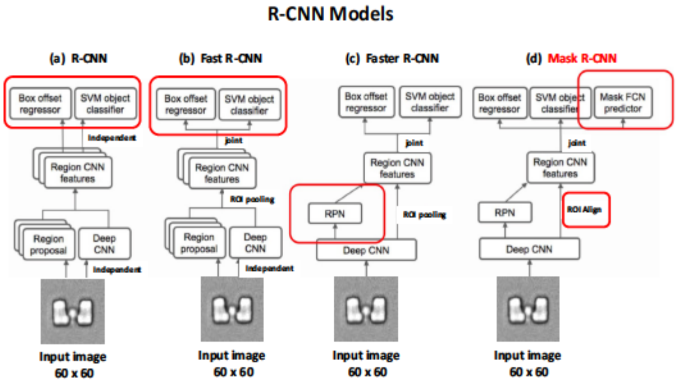

Fig. 3. Development of Region-Based Convolutional Neural Network architectures including (a) R-CNN , (b) Fast R-CNN , (c) Faster R-CNN and (d) Mask R-CNN.

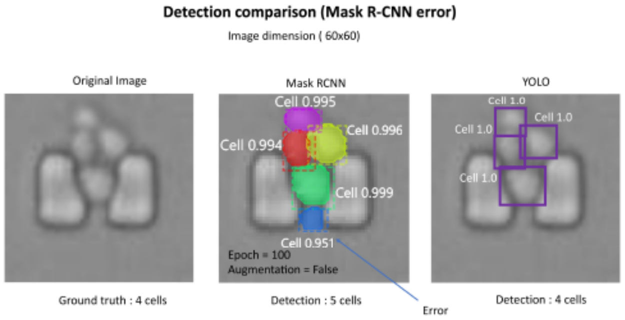

Fig. 4. Model detection comparison for trap with 4 cells. Mask R-CNN detected an extra cell. YOLO detection matched ground truth.

Network (RPN). Fig 3d illustrates Mask R-CNN modifications to Faster R-CNN: ROI pooling is replaced with ROI alignment and a fully convolutional network (FCN) is added to feature analysis for determining object masks.

III. Results and Discussion

A. Cell detection

This study assessed YOLO and Mask R-CNN object detection performance with 60x60 and augmented 512x512 test datasets. We trained YOLO for 200 epochs and Mask R-CNN for 100 and 400 epochs. YOLO performance was evaluated only with dataset augmentation; Mask R-CNN was evaluated both with and without dataset augmentation.

Fig 4 shows detection results for both models. Ground truth was 4 cells, Mask R-CNN detected an extra cell at the trap outlet (blue cell). This illustration is based on 60x60 image dimensions, 200 epochs for YOLO, and 100 epochs for Mask R-CNN without dataset augmentation.

Fig 5 shows Mask R-CNN detecting the correct number of cells, but overestimating cell size. Dataset augmentation enables Mask R-CNN to better estimate cell size.

Fig 6 illustrates dataset augmentation and 400 epochs improving Mask R-CNN cell detection. The detected cell and mask image counts are similar to the source image. In contrast, YOLO detected 1 less cell than ground truth (purple cell).

Fig 7 illustrates the benefit of using larger images created with cubic interpolation. Since YOLO and Mask R-CNN are designed to detect objects at higher image resolution, we supplemented the study with scaled-up images. Fig 7 represents detection accuracy differences due to image size.

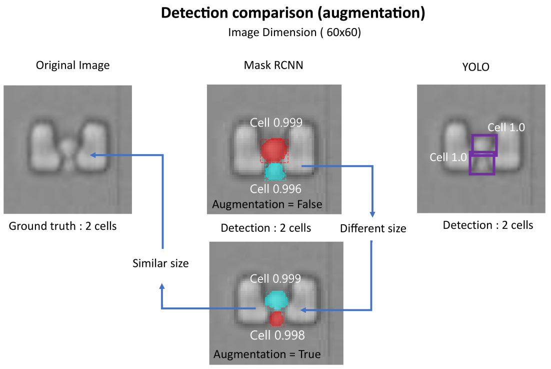

Fig. 5. Detection with dataset augmentation. YOLO and Mask R-CNN detection matched the ground truth. Dataset augmentation improved the accuracy of mask area for Mask R-CNN.

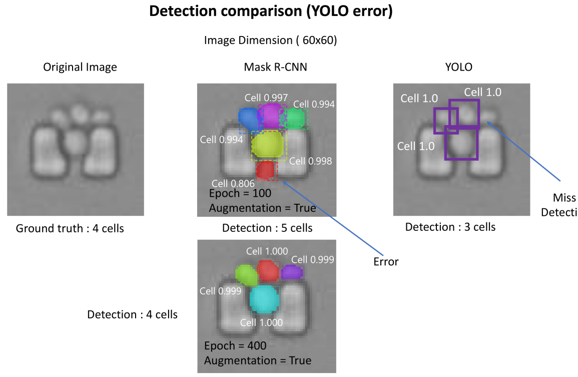

Fig. 6. Modification of Mask R-CNN including YOLO detection error. Mask R-CNN with augmentation and 400 epochs detected 4 cells (matched the ground truth), and YOLO detected 3 cells.

The top-row images show similar results for Mask R-CNN and YOLO with dataset augmentation and 400 epochs applied to Mask R-CNN. The bottom-row shows YOLO detecting 2 cells with an inaccurate bounding box (covering only half the cell area). In this example, Mask R-CNN with data augmentation and 400 epochs was more accurate than YOLO..

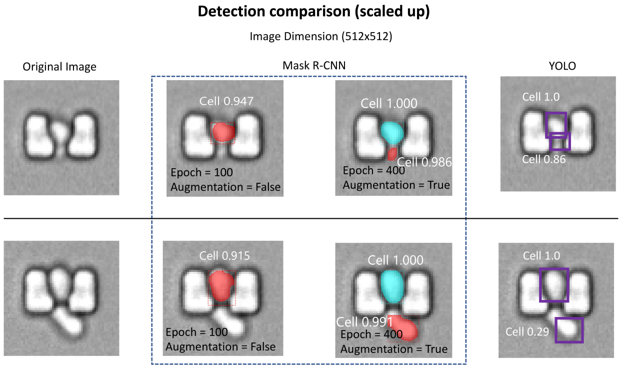

Fig. 7. Comparison of YOLO and Mask R-CNN with higher image resolution. In the first row, modified Mask R-CNN and YOLO matched the ground truth (2 cells). In the second row, YOLO detected a small portion of the cell below the trap.

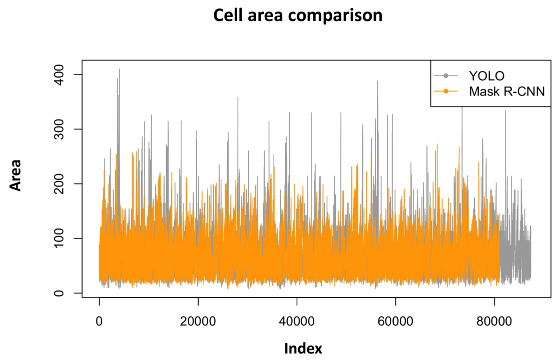

Fig. 8. Cell area comparison for YOLO and Mask R-CNN. Orange/Gray discrepancy illustrates Mask R-CNN detecting fewer cells.

B. Feature extraction

In this section, we evaluate the performance of YOLO and Mask R-CNN on a dataset that contains 100 training images and 40,664 test images. YOLO trained for 200 epochs and Mask R-CNN trained for 400 epochs. Our dataset augmentation was used for both models. Features for both models are 'area', 'total objects', 'confidence', and 'coordinates'.

Fig 8 shows cell size comparison using both models. YOLO results are in gray, Mask R-CNN results are in orange. Yolo's average cell area is larger Mask R-CNN's. YOLO's average cell size ranges from 80 to 100 pixels with confidence rate from 10% to 100%. In contrast, Mask R-CNN's average detected cell size ranges from 50 to 80 pixels, and its detection rate confidence ranges from 90% to 100%.

Fig 9 plots cell size variation versus detection counts for sample traps 01, 20, and 60. for both models. YOLO results show many same-size cells (represented as a row) which indicates that YOLO is less accurate predicting cell size. More variation with Mask R-CNN indicates greater accuracy determining cell size.

Fig 10 shows total counts: 87,908 (YOLO) and 81,842 (Mask R-CNN).

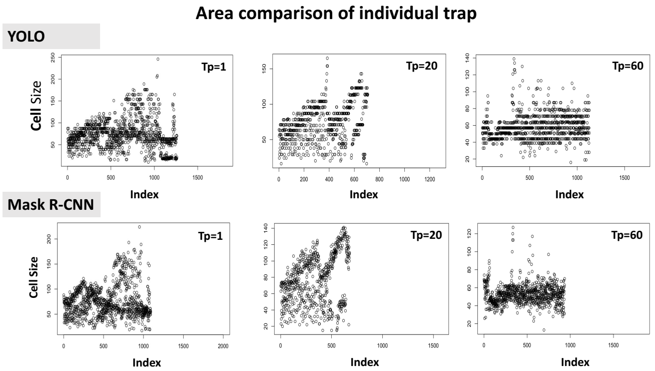

Fig. 9. Area variation for sample traps. YOLO is more accurate for larger cell sizes and Mask R-CNN is more accurate for smaller cell sizes.

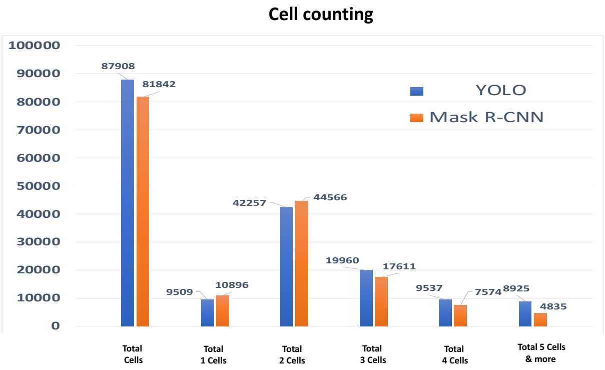

Fig. 10. Cell counting comparison for the individual model. YOLO had better cell detection when the number of cells inside a trap was more the 2 cells. Mask R-CNN performed better when the number of cells was in the range of 1 to 2 cells.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

C. Models comparison

Performance metrics are calculated using equations 1, 2 and 3 where TP is true positives, TN is true negatives, FP is false positives, and FN is false negatives.

Table I compares simple metrics for both methods. The metrics for TP and FP indicate YOLO is more accurate for cell detection with mAPs of 90.6% (YOLO) and 73% (Mask R-CNN). Total cell detections indicate that YOLO is more sensitive for object detection and has less variation in the cell area.

Fig 11 compares mean average precisions (mAPs) for the dataset comprising the first 30 images, indicating YOLO fluctuates less than Mask R-CNN. YOLO cell area calculation uses bounding boxes, decreasing accuracy.

In this work, we modelled cell area as ellipses and calculated it using bounding box information. Both models had the highest performance when there were 2 cells inside traps and had poor performance when there were more than 3 cells inside traps. Mask R-CNN performed much better than YOLO when the number of cells inside the trap is less than 3 cells. Although Mask R-CNN has a lower mAP, its cell area detection is more accurate compared to ground truth. Since Mask R-CNN generates masks, cell area accuracy is much higher than YOLO.

TABLE I
MODELS PERFORMANCE COMPARISON

Metric	**YOLO**	**Mask R-CNN**
	Cell	
TP	74	53
FP	7	11
Precision	95.03%	85%
Recall	92%	82%
Total Detection	81	64
Total Image	30	30
mAP	**90.6 %**	**73 %**

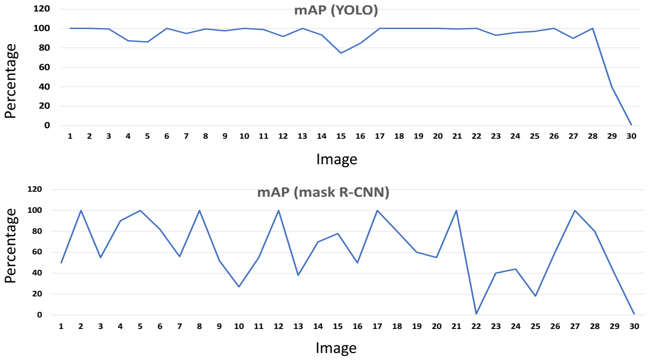

Fig. 11. mAP comparison for YOLO and Mask R-CNN.

IV. Conclusion

We evaluated two CNN models for detecting cells in microfluidic images. YOLO and Mask R-CNN were trained with 100 yeast microfluidic images, tested for object detection on 30 images, and feature extraction on 40,664 images. The results indicate that YOLO was more accurate for object detection but was very sensitive to noise. Yolo also was less accurate for area estimation.

To both generalize and summarize: YOLO appears useful for feature extraction and object detection, but less-so for cell area determination and produces extra unnecessary details (noise). Mask R-CNN produces better estimates of area due to its use of masking and can be improved with data augmentation and increasing epoch count, which increases already computationally expensive training.

This comparison implies that YOLO and Mask R-CNN are both useful for automatic small object detection from medical images. However, we emphasize the present study highlights the need for further development of deep learning methods to facilitate the analysis of time-lapse microscopic images generated by microfluidic devices.

V. Acknowledgement

This work was partially supported by NSF CAREER award #1453078 (transferred to #1720215), NSF award #1761839, a start-up fund, internal awards from the University of Tennessee at Chattanooga. We acknowledge the support of the College of Engineering and Computer Science for a Lambda Quad GPU workstation. We thank the Research as a Service (RaaS) cluster at the SimCenter that supported the prototyping of this work. We also acknowledge the support of NIH grants #R01AG052507 and #R42AG058368.

References

[1] M. McAuliffe, F. Lalonde, D. McGarry, W. Gandler, K. Csaky, and B. Trus, "Medical image processing, analysis and visualization in clinical research," in *Proceedings 14th IEEE Symposium on Computer-Based Medical Systems. CBMS 2001*, 2001.

[2] F. Ritter, T. Boskamp, A. Homeyer, H. Laue, M. Schwier, F. Link, and H.-O. Peitgen, "Medical image analysis," *IEEE Pulse*, vol. 2, no. 6, pp. 60–70, 2011.

[3] M. McAuliffe, F. Lalonde, D. McGarry, W. Gandler, K. Csaky, and B. Trus, "Medical image processing, analysis and visualization in clinical research," in *Proceedings 14th IEEE Symposium on Computer-Based Medical Systems. CBMS 2001*, 2001, pp. 381–386.

[4] M. Ghafari, D. Mailman, and H. Qin, "Application note: polar - an interactive 2d visualization tool for time-series," 2021, pp. 1–6. [Online]. Available: https://ssrn.com/abstract=3827406

[5] R. L. Brocca, F. Menolascina, D. di Bernardo, and C. Sansone., "A novel graphical model approach to segmenting cell images," pp. 131–139, 2012.

[6] M. Kvarnström, K. Logg, A. Diez, K. Bodvard, and M. Käll, "Image analysis algorithms for cell contour recognition in budding yeast," *Opt. Express*, vol. 16, no. 17, pp. 12 943–12 957, Aug 2008. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-16-17-12943

[7] D. Rea, G. Perrino, D. di Bernardo, L. Marcellino, and D. Romano, "A gpu algorithm for tracking yeast cells in phase-contrast microscopy images," *The International Journal of High Performance Computing Applications*, vol. 33, no. 4, pp. 651–659, 2019. [Online]. Available: https://doi.org/10.1177/1094342018801482

[8] S.-C. Chen, T. Zhao, G. J. Gordon, and R. F. Murphy, "A novel graphical model approach to segmenting cell images," in *2006 IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology*, 2006, pp. 1–8.

[9] S.-C. Chen, G. J. Gordon, and R. F. Murphy, "Graphical models for structured classification, with an application to interpreting images of protein subcellular location patterns," *J. Mach. Learn. Res.*, vol. 9, p. 651–682, Jun. 2008.

[10] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 386–397, 2020.

[11] M. A. Al-masni, M. A. Al-antari, J.-M. Park, G. Gi, T.-Y. Kim, P. Rivera, E. Valarezo, M.-T. Choi, S.-M. Han, and T.-S. Kim, "Simultaneous detection and classification of breast masses in digital mammograms via a deep learning yolo-based cad system," *Computer Methods and Programs in Biomedicine*, vol. 157, pp. 85–94, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169260717314980

[12] H. Wu and J. P. Siebert, "Fully convolutional networks for automatically generating image masks to train mask R-CNN," *CoRR*, vol. abs/2003.01383, 2020. [Online]. Available: https://arxiv.org/abs/2003.01383

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[14] S. Y. J. Y. B. Z. S. D. Y. L. L. X. Chenchen Xu, Guili Wang, "Fast vehicle and pedestrian detection using improved mask r-cnn," *Mathematical Problems in Engineering,*, vol. 15, 2020.

[15] A. O. Vuola, S. U. Akram, and J. Kannala, "Mask-rcnn and u-net ensembled for nuclei segmentation," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, 2019, pp. 208–212.

[16] M. C. Jo, W. Liu, L. Gu, W. Dang, and L. Qin, "High-throughput analysis of yeast replicative aging using a microfluidic system," *Proceedings of the National Academy of Sciences*, vol. 112, no. 30, pp. 9364–9369, 2015. [Online]. Available: https://www.pnas.org/content/112/30/9364

[17] P. Salavati and H. M. Mohammadi, "Obstacle detection using googlenet," in *2018 8th International Conference on Computer and Knowledge Engineering (ICCKE)*, 2018, pp. 326–332.

[18] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6517–6525.

[19] ——, "Yolov3: An incremental improvement," *CoRR*, vol. abs/1804.02767, 2018. [Online]. Available: http://arxiv.org/abs/1804.02767

[20] W. Boyuan and W. Muqing, "Study on pedestrian detection based on an improved yolov4 algorithm," in *2020 IEEE 6th International Conference on Computer and Communications (ICCC)*, 2020, pp. 1198–1202.

[21] R. B. Girshick, "Fast R-CNN," *CoRR*, vol. abs/1504.08083, 2015. [Online]. Available: http://arxiv.org/abs/1504.08083

[22] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *CoRR*, vol. abs/1506.01497, 2015. [Online]. Available: http://arxiv.org/abs/1506.01497

[23] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, ser. NIPS'12. Red Hook, NY, USA: Curran Associates Inc., 2012, p. 1097–1105.