MIMO YOLO - A Multiple Input Multiple Output Model for Automatic Cell Counting

1st Hunter Morera
Computer Science and Engineering
University of South Florida
Tampa, USA
hmorera@usf.edu

2nd Palak Dave
Computer Science and Engineering
University of South Florida
Tampa, USA
palakdave@usf.edu

3rd Saeed Alahmari
Computer Science
Najran University
Najran, KSA
saeed3@usf.edu

4th Yaroslav Kolinko
Histology and Embryology
Charles University
Pilsen, CZ
kolinko.yaroslav@gmail.com

5th Lawrence O. Hall
Computer Science and Engineering
University of South Florida
Tampa, USA
lohall@usf.edu

6th Dmitry Goldgof
Computer Science and Engineering
University of South Florida
Tampa, USA
goldgof@usf.edu

7th Peter R. Mouton
SRC Biosciences
Tampa, USA
peter@disector.com

***Abstract*—Across basic research studies, cell counting requires significant human time and expertise. Trained experts use thin focal plane scanning to count (click) cells in stained biological tissue. This computer-assisted process (optical disector) requires a well-trained human to select a unique best z-plane of focus for counting cells of interest. Though accurate, this approach typically requires an hour per case and is prone to inter- and intra-rater errors. Our group has previously proposed deep learning (DL)-based methods to automate these counts using cell segmentation at high magnification. Here we propose a novel You Only Look Once (YOLO) model that performs cell detection on multi-channel z-plane images (disector stack). This automated Multiple Input Multiple Output (MIMO) version of the optical disector method uses an entire z-stack of microscopy images as its input, and outputs cell detections (counts) with a bounding box of each cell and class corresponding to the z-plane where the cell appears in best focus. Compared to the previous segmentation methods, the proposed method does not require time- and labor-intensive ground truth segmentation masks for training, while producing comparable accuracy to current segmentation-based automatic counts. The MIMO-YOLO method was evaluated on systematic-random samples of NeuN-stained tissue sections through the neocortex of mouse brains (n=7). Using a cross validation scheme, this method showed the ability to correctly count total neuron numbers with accuracy close to human experts and with 100% repeatability (Test-Retest).**

***Index Terms*—Deep Learning, Neuron Counts, YOLO, Microscopy, Stereology**

I. Introduction

Changes in the number of specific cells on histological sections is an important metric in many fields of biomedical research involving cell degeneration, cytotoxicology, and cellular inflammation. State-of-the-art unbiased stereology methods to make accurate cell counts in tissue sections require trained humans to focus through a z-stack of microscopy images and manually count (click) on hundreds of cells per case. This manual counting is both time consuming and prone to human error due to subjective decision-making about the cells counted.

There is a strong interest in the development of accurate and efficient methods for automatic counts of specific cells in microscopy images through known tissue volumes [1]–[7], such as immunohistochemical (IHC) stained images obtained by ordinary bright-field microscopy. Due to challenges of three-dimensional (3D) cell counts, work to date has focused on Extended Depth of Field (EDF) images for automatic cell counting [8], [9]. However, in tissues with high cell packing densities, cell counts on EDF images can have variable levels of under-counts due to overlapping and masking of cells at different z-plane locations.

More recent work has focused on deep learning (DL)-based segmentation methods for automation of cell counts using unbiased methods. One such method uses a similar multiple input multiple output (MIMO) framework to segment cells in a z-axis stack of images (disector stack) where every cell is segmented in its best focus z-plane. This MIMO U-Net approach [10] shows high accuracy ($< 10\%$ error) on NeuN-immunostained neurons sampled in an unbiased manner. A drawback of this approach is the requirement for manual annotation of ground truth (GT) masks for training the DL model. These masks are exact outlines of neurons at their plane of best focus in z-stack images. The need to overcome this drawback has led us to explore cell detection, rather than cell segmentation, which only requires GT bounding boxes of each cell as opposed to more time- and labor-intensive fine outlines of cells, nuclei, cell bodies, etc.

One of the most popular object detection architectures is You Only Look Once (YOLO) introduced in 2015 [11]. Advantages of this framework are its balance between accuracy and speed of detection. Across its iterations, YOLO has proved useful in many domains including agriculture [12] and medicine [13]. Due to the limitations of bright-field IHC data described above, 3D processing is not practical and the z-

Funding provided by the National Science Foundation and Florida High Tech Corridor.

stacks are treated as sequential data. In this paper, we describe our modifications to YOLOv5 [14] to use as input z-stacks in the form of multi-channel images. The goal is to output bounding boxes for every cell (X and Y location) along with a class prediction that corresponds to an individual z-plane for each counted cell.

To evaluate the performance of our MIMO YOLO approach, we have used a dataset consisting of z-stack images from neurons stained by NeuN IHC. Sections and z-axis image stacks sampled using unbiased methods were obtained from the neocortex (NCTX) of the mice brains and manually counted at 100x magnification by trained experts. The same images were then used in a cross-validation approach where the model was trained on six mice and tested on one. Results showed the model was able to correctly count cells with less than 10% average error ($>$ 90% accuracy) when compared to human experts. Our MIMO YOLO detection approach showed no statistical difference in cell counting results when compared to previous MIMO U-Net results on the same z-stacks, while MIMO YOLO reduced the training time by half when compared to MIMO U-Net. This novel approach showed 100% repeatability (Test-Retest) for each case. Using the trained MIMO YOLO, time was reduced from about one hour for manual cell counts to approximately 30 seconds for automatic counts per case.

II. Data

The dataset used as described in [10] consists of bright-field microscopy images collected from the NCTX in seven mice brains. Systematic-random sets of sections through each case were stained with IHC for NeuN neurons. Images were collected at 100x magnification using an Olympus BH-2 microscope with a 100x oil lens and motorized XYZ stage and Neuron counts were done by a trained expert (YK) using the manual optical disector method. For each mouse, approximately 60 image stacks on average were collected with five images per stack using a z-axis separation of two microns per image; therefore, each sampled disector stack included NeuN neurons in 10-μm tissue volume. These color images were first converted to grayscale using a correlation-based method [15]. Each of the five-plane z-stacks images were 256x256 pixels, where GT is a bounding box for each cell counted by a human expert. The class of each bounding box (object) is equal to the best focus z-plane for that cell. An example image stack can be seen in Fig. 1.

III. Method

Object detection networks are deep neural networks trained to provide the location and class of an object in an image. One of the most popular of these object detection networks is YOLO [11]. Normally, YOLO takes in a three-channel RGB image and outputs (detects) objects in the image with bounding box coordinates and a class of the object present in that box. In this section, we describe our changes to YOLO for multi-channel input cell detection.

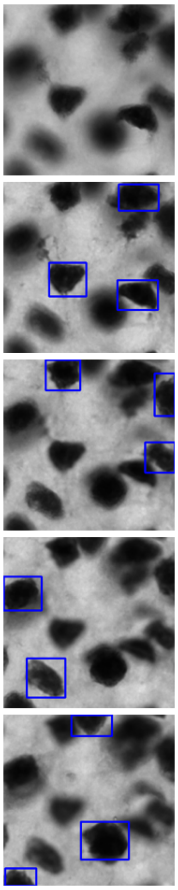

Fig. 1. Example z-stack of images with GT bounding boxes in blue.

A. MIMO YOLO

The MIMO YOLO method uses a z-stack of images as input to the neural network and outputs a bounding box for every cell in a z-stack, with the class of the bounding box (object) corresponding to the image in the z-stack where each cell is in best focus. To accomplish this goal, we first pass an entire z-stack into the model to provide bi-direction context to determine the plane of best focus. This process begins by converting the z-stack images from RGB to grayscale using a correlation-based method [15]. This step provides single channel images for each of the images in the z-stack. The images from a single z-stack can then be combined into a single multi-channel Tiff image. The number of channels will depend on the number of images in the z-stack. The data used in this paper resulted in a 5-channel image, where each channel is a single channel z-stack image.

As available on GitHub, YOLOv5 [14] was designed to work with only 3-channel images. Changes needed to be made to the data-loader and various data processing functions to allow the code to work with any number of channels. The code also comes with built in on-fly augmentation which is applied to image batches prior to training. In some cases, like augmenting hue, saturation, and value (HSV), they needed to be removed, as we were not using color images. The rest of

the built-in augmentation worked with multi-channel images. The GT for the training images were generated by placing a bounding box around each cell counted by an expert. The class for each box was equal to where in the z-stack the cell was in best focus. Thus, for the data used here we would have the option of class 0, 1, 2, 3, or 4 for a bounding box. The zero class corresponded to best focus in the first channel of the five channel image, and class of four would correspond to best focus in the last channel of the five channel image.

The models were trained using the built-in learning rate scheduler along with the Stochastic Gradient Descent (SGD) optimizer. Training was done for 500 epochs on every fold and the pre-trained common objects in common context (COCO) dataset weights were used instead of random weight generation to start. The images were augmented using on-fly augmentation of random rotation, translation, scaling, sheering, flipping, and mosaic. In addition to these on-fly augmentations, images were also pre-augmented using z-axis flipping of the z-stacks, doubling the size of original data. To test the performance of the model, we split the data into seven folds where each had one mouse left out for unseen test data. The rest of the data from six mice were split 80% for training and 20% for validation. This resulted in an average training set size of 609 images per fold, and 79 validation images per fold. Due to these small training sets, on-fly augmentation, and the speed of YOLOv5, we were able to train a full model in around 25 minutes and predict on the testing data and apply post-processing in less than 30 seconds per case.

B. Post-processing

Once the model was trained, predictions were made on the test data. The model used for these predictions was the one with the best precision and recall on the validation set. The inference code from YOLOv5 uses non-maximal suppression (NMS) to remove extraneous bounding boxes. This function takes parameters of intersection over union (IOU) and confidence threshold, which it uses to remove bounding boxes from the final output. For the results shown, we are using a confidence threshold of 0.40 and IOU threshold of 0.45 as these produced best performance on validation data. However, in addition to NMS applied by the YOLOv5 code, we apply a second level of post-processing to the predictions before evaluation. An example of the workflow can be seen in Fig. 2

The most common mistake made by the network in predictions is having multiple bounding boxes per cell with different classes corresponding to best focus. For example, if a cell's best focus is in plane 3, we often see a bounding box with class 3 and one with either class 4 or class 2. We call these duplicate z-plane detections. To combat this issue, our second level of post-processing involves removing these extra detections when they occur. We do this by evaluating all bounding boxes for duplicate z-plane detections with at least a 0.60 IOU with another from a z-plane directly above or below. The bounding box with the highest confidence as provided by the model is kept, and the others discarded. This approach ensures accurate count in the z-stack since each cell is detected and counted only once in its best focus plane.

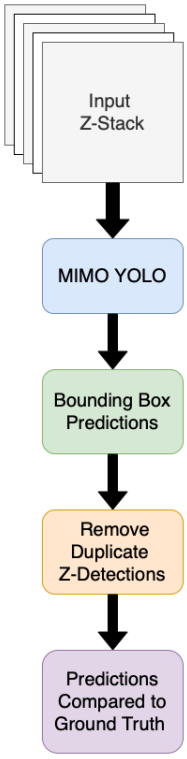

Fig. 2. MIMO YOLO work flow.

C. Evaluation

After the model prediction and post processing is done on the test images, the performance is evaluated. A cell is considered correctly counted if there exists a predicted bounding box overlapping the ground truth bounding box with an IOU of at least 0.60, and it is within one z-plane of the expert's best focus plane. This buffer of one z-plane is given because experts determine which of the planes the cell is in best focus, and in some cases, this can be hard to distinguish. If no prediction exists that meets this criteria for a cell counted by the expert, it is considered missed or as a false negative. Those predictions which are not considered correctly counted cells make up the false positive count. The main metrics relied on for model evaluation are count error rate and F1 score. Count error rate calculation can be seen in (1) which uses GT cell count (GT_Count) and predicted cell count (Pred_Count). The F1 score calculation can be seen in (2), and uses true positive count divided by predicted count for precision, and true positive count divided by GT count for recall.

$$Count\ Error\ Rate = \frac{|GT\_Count - Pred\_Count|}{GT\_Count} \quad (1)$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \quad (2)$$

IV. Results

Each model was evaluated on the test animal that was left out during training. The predicted labels were processed and evaluated as described above. The results showed an average count error rate of 7.34% for the seven folds. The results for MIMO YOLO can be seen in Table I and results for MIMO U-Net in Table II. For reference, manual inter-rater count error is, on average, about 5%. As stated, the F1 score is most important here, as it supplies information about the balance of false positives and false negatives. When it comes to overall count, these can have a canceling effect which results in the count error rate being low, despite the model making some errors. A student's t-test was done with a hypothesis of a statistical difference with $p<0.05$ to compare MIMO YOLO (detection) and our previous MIMO U-Net (segmentation) model. The results of the student's t-test on the same folds (i.e., same training and testing data) showed no statistical difference between the count errors with $p = 0.80$. An example of the final output can be seen in Fig. 3, where the green bounding boxes are for GT and the red are for predictions. An example of a missed cell or false negative can be seen in Fig. 3 on the last image in the stack. Despite these types of errors, this example demonstrates the ability of our MIMO YOLO method to learn both X and Y location of cells and the best focus plane.

TABLE I
MIMO YOLO Average Results of Seven Folds

	Count Error (%)	**Precision**	**Recall**	**F1-Score**
Mean	7.34	0.82	0.80	0.81
STD	5.70	0.04	0.05	0.02

TABLE II
MIMO U-Net Average Results of Seven Folds

	Count Error (%)	**Precision**	**Recall**	**F1-Score**
Mean	6.56	0.80	0.82	0.81
STD	5.75	0.06	0.04	0.04

V. Conclusion

The time currently required for a human expert to make unbiased cell counts in microscopy images (∼1 hour per case) is a limitation in many biomedical research fields. In this work, we present MIMO YOLO, a multiple input multiple output version of YOLO which can be used to count cells in bright field microscopy images. IHC z-stacks of images are impractical to process in 3D due to out-of-focus structures and relatively low SNR when compared to immunofluorescent microscopy images. Thus, z-axis stacks must be processed as sequence data to make accurate counts of cells contained within those z-stacks. We described the modifications necessary to YOLOv5 to use z-stacks as multi-channel input images.

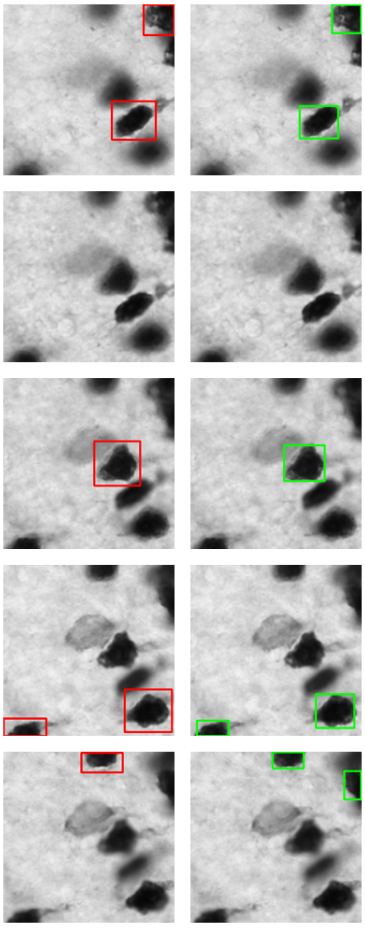

Fig. 3. Left column shows predicted bounding boxes in red. Right column shows ground truth bounding boxes in green.

The GT for these multi-channel microscopy images are bounding boxes for each cell counted by a trained expert, with the class of that bounding box (object) equal to the plane which the cell is in best focus. The proposed method is intended to overcome the time, labor, and potential for low inter-rater error inherent to subjective human counts. This subjectivity is reduced here through the use of GT from a well-trained human expert. The proposed MIMO YOLO method reduces the time needed to perform cell counts by close to 99% compared to human experts. Using a professionally collected and annotated dataset from seven mice of NeuN stained neurons, we showed counts with less than 10% error when compared to trained experts. Furthermore, MIMO YOLO provides counts with no statistical difference to cell segmentation by MIMO U-Net. In terms of the effort required for building DL models, our proposed cell detection approach requires GT as a simple bounding box after expert counts, as opposed to tedious cell outlines around each cell required for cell segmentation by MIMO U-Net. Though both approaches still require a human expert, the use of a bounding box substantially reduces the demands on the expert's time and effort. Our MIMO YOLO approach also reduces the time needed to train a model by

half when compared to MIMO U-Net on the same data.

VI. Future Work

The present work focused on automatic counts of NeuN-immunostained neurons in the NCTX of mice brains. Our future work includes expanding the MIMO YOLO approach to other datasets with different biostructures stained by different staining methods, for example, immunofluorescent Rab-5 endosomes on confocal microscopy images from mouse brains. Other work will include ensembles of MIMO YOLO networks, various parameter tuning, and post-processing methods to further reduce count error rates to 5% or below compared to human experts. Finally, we will explore combinations of MIMO YOLO and MIMO U-Net as a possible approach to increase performance beyond that of individual methods.

Compliance with Ethical Standards

The use of animals in this work complies with federal regulations regarding the care and use of laboratory animals: Public Law 99-158, the Health Research Extension Act, and Public Law 99-198, the Animal Welfare Act which is regulated by USDA, APHIS, CFR, Title 9, Parts 1, 2, and 3.

Acknowledgment & Funding

This work was supported by National Science Foundation Grants (#1513126, #1746511, #1926990) and a Florida High Tech Corridor Grant (ENG203, #20-10) to SRC Biosciences and the University of South Florida. Yaroslav Kolinko was supported by the Cooperation Program (research area MED/DIAG).

References

[1] M. Oberlaender, V. J. Dercksen, R. Egger, M. Gensel, B. Sakmann, and H.-C. Hege, "Automated three-dimensional detection and counting of neuron somata," *Journal of neuroscience methods*, vol. 180, no. 1, pp. 147–160, 2009.

[2] J. D. Ross, D. K. Cullen, J. P. Harris, M. C. LaPlaca, and S. P. DeWeerth, "A three-dimensional image processing program for accurate, rapid, and semi-automated segmentation of neuronal somata with dense neurite outgrowth," *Frontiers in neuroanatomy*, vol. 9, p. 87, 2015.

[3] B. Mathew, A. Schmitz, S. Muñoz-Descalzo, N. Ansari, F. Pampaloni, E. H. Stelzer, and S. C. Fischer, "Robust and automated three-dimensional segmentation of densely packed cell nuclei in different biological specimens with lines-of-sight decomposition," *BMC bioinformatics*, vol. 16, no. 1, pp. 1–14, 2015.

[4] G. Mazzamuto, I. Costantini, M. Neri, M. Roffilli, L. Silvestri, and F. S. Pavone, "Automatic segmentation of neurons in 3d samples of human brain cortex," in *Applications of Evolutionary Computation: 21st International Conference, EvoApplications 2018, Parma, Italy, April 4-6, 2018, Proceedings 21*, pp. 78–85, Springer, 2018.

[5] R. Li, M. Zhu, J. Li, M. S. Bienkowski, N. N. Foster, H. Xu, T. Ard, I. Bowman, C. Zhou, M. B. Veldman, *et al.*, "Precise segmentation of densely interweaving neuron clusters using g-cut," *Nature communications*, vol. 10, no. 1, p. 1549, 2019.

[6] H. Wang, D. Zhang, Y. Song, S. Liu, Y. Wang, D. Feng, H. Peng, and W. Cai, "Segmenting neuronal structure in 3d optical microscope images via knowledge distillation with teacher-student network," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pp. 228–231, IEEE, 2019.

[7] A. LaTorre, L. Alonso-Nanclares, J. M. Peña, and J. DeFelipe, "3d segmentation of neuronal nuclei and cell-type identification using multichannel information," *Expert Systems with Applications*, vol. 183, p. 115443, 2021.

[8] S. Alahmari, D. Goldgof, L. Hall, P. Dave, H. A. Phoulady, and P. Mouton, "Iterative deep learning based unbiased stereology with human-in-the-loop," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 665–670, IEEE, 2018.

[9] S. S. Alahmari, D. Goldgof, L. Hall, H. A. Phoulady, R. H. Patel, and P. R. Mouton, "Automated cell counts on tissue sections by deep learning and unbiased stereology," *Journal of chemical neuroanatomy*, vol. 96, pp. 94–101, 2019.

[10] P. Dave, Y. Kolinko, H. Morera, K. Allen, S. Alahmari, D. Goldgof, H. O. Lawrence, and P. R. Mouton, "Mimo u-net: efficient cell segmentation and counting in microscopy image sequences," in *Medical Imaging 2023: Digital and Computational Pathology*, SPIE, 2023.

[11] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.

[12] Y. Tian, G. Yang, Z. Wang, H. Wang, E. Li, and Z. Liang, "Apple detection during different growth stages in orchards using the improved yolo-v3 model," *Computers and electronics in agriculture*, vol. 157, pp. 417–426, 2019.

[13] J. George, S. Skaria, V. Varun, *et al.*, "Using yolo based deep learning network for real time detection and localization of lung nodules from low dose ct scans," in *Medical Imaging 2018: Computer-Aided Diagnosis*, vol. 10575, pp. 347–355, SPIE, 2018.

[14] G. Jocher, Ayush Chaurasia, A. Stoken, J. Borovec, NanoCode012, Yonghye Kwon, Kalen Michael, TaoXie, Jiacong Fang, Imyhxy, , Lorna, Zeng Yifu, C. Wong, Abhiram V, D. Montes, Zhiqiang Wang, C. Fati, Jebastin Nadar, Laughing, UnglvKitDe, V. Sonck, Tkianai, YxNONG, P. Skalski, A. Hogan, Dhruv Nair, M. Strobel, and M. Jain, "ultralytics/yolov5: v7.0 - yolov5 sota realtime instance segmentation," 2022.

[15] H. Z. Nafchi, A. Shahkolaei, R. Hedjam, and M. Cheriet, "Corrc2g: Color to gray conversion by correlation," *IEEE Signal Processing Letters*, vol. 24, no. 11, pp. 1651–1655, 2017.