MIMO U-Net: Efficient Cell Segmentation and Counting in Microscopy Image Sequences

Palak Dave[a], Yaroslav Kolinko[b], Hunter Morera[a], Kurtis Allen[a], Saeed Alahmari[c], Dmitry Goldgof[a], Lawrence O. Hall[a], and Peter R. Mouton[a,d]

[a]Department of Computer Science and Engineering, University of South Florida, Tampa, Florida, 33620, USA
[b]Department of Histology and Embryology & Biomedical Center, Faculty of Medicine in Pilsen, Charles University, Pilsen, Czech Republic
[c]Department of Computer Science, Najran University, Najran, 66462, KSA
[d]SRC Biosciences, Tampa, Florida 33606, USA

1. INTRODUCTION

Quantification of stained cells in microscope images is important in many fields of biomedical research. In bright-field microscopy, a tissue slide is homogeneously illuminated as opposed to confocal and multi-photon microscopy where a precisely focused laser beam spatially filtered through a pin-hole is used. Hence, confocal and multi-photon microscopy are capable of providing a signal from only one thin focal plane. Such z-stacks allow for 3D segmentation/reconstruction of cell structures due to no out-of-focus structures.[1] Moreover, the fluorescent dyes used in widefield, confocal, and multi-photon microscopy also contribute to a high Signal-to-Noise Ratio (SNR) compared to the immunohistochemical staining used in brightfield microscopy. Various methods are available for 3D analysis of neurons in fluorescent modalities.[2–8] However, such a method for brightfield microscopy z-stacks is a challenge because of the presence of background signal from out-of-focus structures and low SNR.[1] Štajduhar et al. in[9] performed NeuN neuron counting on a single image collected at low magnification (40x) resulting in thicker focal plane. Another work in[1] localized neurons in the z-direction based on the image sharpness of a neuron body. The neurons are first detected on only the mid-plane of the z-stack using the method from.[9] However, highly overlapping or masked neurons at different z-depths can result in under-counting. A similar issue occurs in another approach where neuron counting is done on an Extended Depth of Field (EDF) image.[10] Our recent publication[11] for neuroscience audience has shown that accurate counting of overlapping or masked cells can be achieved by segmenting each cell in its best focus image, to leverage the z-separation between the cells, using a MIMO approach. In this manuscript, we present algorithmic details of the MIMO approach, report results on a new dataset, and release the dataset and code. Also, the MIMO approach is applied to a publicly available dataset with a suitable configuration and compared against a more sophisticated method.

A z-stack can be "viewed" as either volumetric or sequential data. 3D segmentation is impractical for our data because of poor SNR and signals from out-of-focus neurons as described above. Hence, each z-stack is treated as a set of sequential images in the present work.

The neuron segmentation in its best focus plane is a binary segmentation task where the foreground for a z-image consists of neurons in the best focus in the given z-image. We refer to this task as best-focus-neuron segmentation task. Identification of the best focus image strongly requires bi-directional context from the images above and below the target z-image. Hence, it is an intuitive approach to use sequence processing methods such as bi-directional Recurrent Neural Networks (RNNs). However, the computational cost of such methods is high due to the high neural network parameters, longer training time and requires larger training data which is often a limiting factor in biomedical applications. Moreover, parallelization of such a network is also challenging.

There is a line of research that treats a z-stack as multiple input channels to a 2D Fully Convolutional Network (FCN) thereby leveraging the sequential information and each of the multiple output channels represents one of the multiple output classes (Multiple Input Multiple Output - MIMO).[12] We propose to first, pose the two-class best-focus-neuron segmentation problem as a multi-class multi-label problem by considering foreground in n^{th} z-image as output class "n" and second, utilize the channel axis of a 2D U-Net in a MIMO setup to obtain bidirectional sequential context. This approach is referred to as MIMO U-Net hereafter.

The proposed MIMO U-Net approach is also applied to image sequences of a different nature, fluorescent time-lapse microscopy. The reason behind the application of MIMO U-Net for time-lapse microscopy is not to resolve overlapping cells using z-separation but to validate its sequential context learning capability since the time-lapse microscopy image sequences are 2D+t in contrast to the 2D+z z-stacks. Here, a cell has to be segmented in every

time-stamp t-image instead of only the best focus z-image. We used the Fluo-N2DH-GOWT1 dataset from the Cell Tracking Challenge - CTC (http://celltrackingchallenge.net/) and the result is compared against a U-LSTM method designed for the CTC datasets.[13] It is important to note that the present work does not claim the best result on the CTC dataset. The sole purpose of this activity was to show that the proposed MIMO approach using a 2D U-Net for exploiting the sequential features performs equally well as the U-LSTM method where C-LSTM memory units are integrated into a 2D U-Net architecture to leverage the sequential context on a publicly available benchmark dataset.

There are several advantages of the proposed MIMO U-Net method. First, a minimal increase in network parameters compared to vanilla U-Net. Second, high performance with a small amount of training data as compared to other sequence processing models due to a smaller number of trainable parameters. Third, MIMO U-Net can provide bidirectional context by default compared to other methods where additional network complexity is required to obtain bidirectional context. Fourth, the MIMO formulation can process multiple images as one sample vs one image as one sample, leading to a significant reduction in training and inference time. Finally, it allows for a larger input size, thereby making more context available for learning.

2. METHODS

2.1 MIMO U-Net

A 2D U-Net model consisting of four down- and up-sampling layers with multiple input channels and multiple output classes was used for the multi-class multi-label segmentation. In the proposed MIMO approach, each image of an image sequence is treated as an input channel and the foreground for the n^{th} input channel is considered as the n^{th} output class. The intuition behind using 2D U-Net for image sequence analysis is that any FCN can learn sequential information across input channels since the convolution kernel depth in the first layer is equal to the number of input channels and equal to the number of feature map channels, from the previous layer, in the second layer and so on. Also, each output class is correlated to the other classes. The MIMO approach provides for a computationally efficient mutualism where each image in an input sequence receives context from the other images and provides context to the other images at the same time.

Best-focus-neuron segmentation: There is a strong inter-dependence for the most part among output class probabilities. Hence, the softmax function was used in the last layer for classification. Softmax assumes the target probability distribution to sum up to one. The overlapping cell regions belong to multiple classes - a class for each of the overlapping cells. The label vector is converted to shared-one-hot where each non-zero entry is $1/k$ for $k \geq 1$ classes for the XY-location.[14] A background class is added as one of the output classes to obtain a cleaner softmax distribution in the output. The XY-locations of the z-stack not belonging to any of the foreground classes belong to the background class. Hence, the number of output classes is one more than the number of inputs in this segmentation task as shown in Fig. 1. The loss is computed as $\sum_{i=1}^{c} class\_weight_i * binary\_crossentropy(y\_pred_i, y\_true_i)$ where c is number of output classes and class weights are used to balance the output classes.

CTC dataset: In contrast to the best-focus-neuron counting (2D+z), a cell should be segmented in every time instance t-image (2D+t). The output class probabilities are independent of each other. Hence, Sigmoid activation was used in the last layer. The same loss function as used in best-focus-neuron segmentation was used with the class weights vector set to 1 since the output classes in this task are not imbalanced. As shown in Fig. 2, there is no background class in the output classes since sigmoid activation can predict none of the output classes for a background pixel. A sequence of 92 images was divided into multiple overlapping sub-sequences of 10 images with a stride of 1. Each sub-sequence was used as an input to the MIMO U-Net.

2.2 Post-processing

Best-focus-neuron segmentation: Post-processing, tuned on a random subset of thirteen stacks from the dataset, was performed to prune over-segmentations and False Positives. A minimum area threshold of 500 pixels (which is close to the smallest Ground Truth (GT) blob in the dataset) was used for the blobs completely inside the image and 200 pixels was used for the blobs touching the edges of the image since it can be a cell partially outside the field of view. The z-images in our dataset are 2 microns apart and the neurons are 3D spherical structures (typically > 4 microns in diameter). Hence, it is very likely that any two overlapping blobs

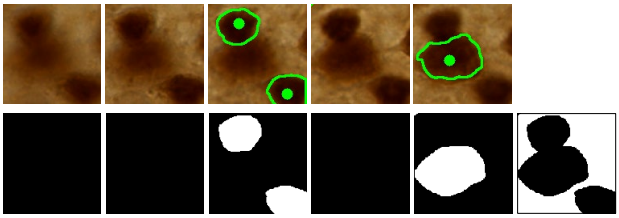

Figure 1: Top row: Segmentation results (green contours) overlaid on crops of the z-images of a stack using the MIMO U-Net. Green dots indicate GT. Bottom row: GT masks for the 5 foreground and 1 background as output classes.

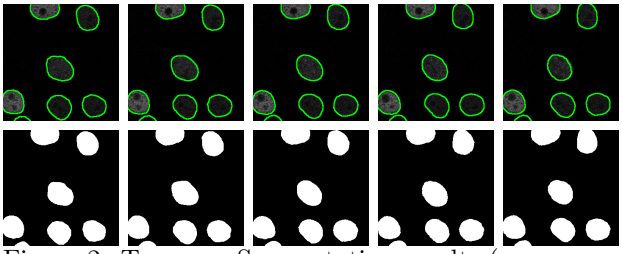

Figure 2: Top row: Segmentation results (green contours) overlaid on crops of five consecutive t-images of a test sequence from CTC dataset. Bottom row: GT masks for the 5 foreground output classes. No background class is required as discussed in Section 2

in any two consecutive z-images belong to a single cell. such blobs are combined and assigned to the z-image with the bigger blob. Hole filling was performed to account for a solid cell body. Finally, blobs having less than a predefined overlap (here, 30%) with a minimum enclosing circle are discarded since neurons are spherical for the most part. **CTC dataset:** For evaluation on the challenge set, the multiple prediction maps for an image resulting from overlapping sub-sequences were averaged to obtain a single prediction map. Blobs having area smaller than the minimum blob area for a cell blob in the train set (438 pixels) were discarded. After hole filling, distance transform seeded watershed segmentation was applied to the blobs larger than double the average area of cells (3000 pixels) in the training data.

3. DATASET AND EVALUATION METRICS

Dataset: The best-focus-neuron segmentation dataset consists of brightfield microscopy z-stacks, captured using an Olympus microscope with a 100x oil lens, from tissue sections of the neocortex region of seven mouse brains. The tissue sections are stained with DAB immunostain for NeuN neurons. This dataset was locally collected and labeled by an expert. There are $\sim$ 60 z-stacks per mouse. Each z-stack has five images ($2\mu m$ apart). The color images were converted to gray-scale using RGB channel weights obtained by applying a correlation-based method from[15] to a random subset of 28 images. The fluorescent time-lapse microscopy image dataset from the Cell Tracking Challenge, namely Fluo-N2DH-GOWT1, has two sequences of 92 8-bit gray-scale images in each of the train and test sets.

Evaluation Metrics: While evaluating on the z-stacks, the objective is to count the neurons using the unbiased stereology rule.[11] Hence, accuracy, precision, recall, and F1-score at an object (neuron) level are used as evaluation metrics. A blob in GT and a predicted blob are called a match if the GT blob's centroid is inside or within a 10-pixel distance from the closest boundary of the predicted blob. The prediction blob is not restricted to be in the same z-image as the GT blob. The matching is one-to-one to ascertain that no predicted blob is matched with more than one highly overlapping or obscured GT blobs. For the CTC dataset, a submitted result on the test set was evaluated by the challenge organizers for detection accuracy (DET), segmentation accuracy (SEG), and overall performance in cell segmentation benchmark (OP_{CSB}).[16,17]

4. EXPERIMENTS AND RESULTS

Best-focus-neuron segmentation: Given the dataset of z-stacks from seven mice, Leave-One-Out (LOO) cross-validation experiment was performed. A random split of 80:20 of the stacks from each of the six mice contributed to the training and validation sets, respectively. Adam optimizer with learning rate=$1e^{-4}$, first moment coefficient $\beta 1$=0.9, second moment coefficient $\beta 2$=0.999, $eps = 1e^{-7}$, and weight decay=$1e^{-3}$ was used.[18] With a batch size of 16, and standardization for feature scaling the model was trained for 100 epochs and the model at the epoch with the smallest validation loss was used for evaluation. The prediction confidence threshold for each fold was selected based on the validation set. The input images of size 256×256 were augmented using x, y, and z-flip. The LOO results are reported in Table 1. The training time per fold was $\sim$ 35 mins on NVIDIA GeForce GTX 1080Ti GPU with Cuda 10.2. The segmentation result is visualized in Fig. 1.

CTC Dataset: After applying histogram equalization to each image for contrast enhancement, overlapping sub-sequences of 10 images with a stride of 1 were generated from each of the two training sequences. A random

	Count Error (%)	Accuracy	Precision	Recall	F1-Score
Mean	6.56	0.68	0.80	0.82	0.81
STD	5.75	0.05	0.06	0.04	0.04

Table 1: Average Neuron counting result over seven folds of leave-one-mouse-out cross validation experiment using the proposed method.

Method	DET	SEG	OP_{CSB}
MIMO U-Net	0.924	**0.883**	**0.903**
U-LSTM	**0.937**	0.854	0.896

Table 2: Evaluation on CTC dataset Fluo-N2DH-GOWT1.

split of 80% and 20% into the sub-sequences from each sequence contributed to the train and validation set, respectively. Input size 1024×1024 was the same as the original image and batch size was 2. The prediction confidence threshold was set to 50%. Other training aspects were kept the same as described above for best-focus-neuron segmentation. The evaluation of the submitted result is reported and compared against the U-LSTM method in Table 2. The SEG result was reported in the U-LSTM paper,[13] and corresponding DET and OP_{CSB} values are obtained from the CTC website.

5. DISCUSSION AND CONCLUSION

Manual labeling for neuron counting is a subjective task. About 3%-5% inter-rater variability in neuron counts has been observed.[19,20] Along with indicating a lower bound for count error, the inter-rater variability also suggests limitations in terms of recall and precision (depending on common and exclusive neurons among raters).

The proposed method is not compared against single z-plane input (vanilla) U-Net for best-focus-neuron counting since it is not possible to ascertain the optimal focus without bidirectional sequential context. Although such a network is expected to learn some meaningful features (e.g. the notion of "well-focusedness" and its correlation with the chance of being an optimal focal plane), it results in a very noisy learning environment. Also, we applied the U-LSTM method for the best-focus-neuron segmentation task. As anticipated, the network failed to learn the best focus and segmented a cell in all relatively good focus planes because only one-dimensional context is available from the C-LSTM memory units.

One weakness of the best-focus-neuron counting is the assumption of only one "optimal" focal plane for a neuron. If multiple planes show a similar level of good focus, it may become a source of ambiguity for the model during learning. Such cases in prediction are handled by the post-processing step in which overlapping detections in consecutive slices are merged, allowing for only one detection per cell. Such post-processing can merge two touching cells in consecutive planes. A possible solution for that can be to allow merging only if the resulting blob size is not larger than the largest cell size in the training data. The extent of this issue depends on the step size in the z-stacks with respect to the size of the cells. A limitation of the proposed MIMO approach is that one model per cell type is required if multiple types of cells are required to be segmented in the same images. However, this limitation is not very restrictive because of the small training time and computational cost.

The U-LSTM method has C-LSTM memory blocks integrated in the U-Net to facilitate the sequential information learning.[13] Notably, U-Net alone in the proposed MIMO approach performed equally well on the CTC dataset. The MIMO U-Net requires a smaller training time because first, it has a small increase in the number of trainable parameters as compared to the vanilla U-Net since it only increases the number of parameters in the first and the last layer. The number of trainable parameters for an input size of 256×256 in vanilla U-Net, MIMO U-Net, and U-LSTM is 7760k, 7761k, and 74607k, respectively. And second, the output for multiple z/t-images is obtained at the same time, as opposed to processing each z/t-image as an individual sample. The decent performance on the CTC dataset shows that the proposed MIMO framework can also be used for other sequential image data like 3D segmentation of brain tumors. Also, the proposed method can be applied to large sequences with varying lengths by using the overlapping sub-sequence approach as used on the CTC dataset.

To conclude, we presented an approach for best-focus-neuron segmentation to resolve overlap using z-separation in z-stacks of brightfield microscopy, where 3D segmentation is not feasible due to out-of-focus signals and low SNR. Furthermore, we propose to utilize a 2D U-Net with a MIMO formulation for inter-image feature learning in microscopy image sequences by posing the binary segmentation problem as a multi-class, multi-label problem. Its advantages include less trainable parameters, small training time, less training data requirement, and availability of bi-directional context without additional neural network complexity. The proposed method achieved an average neuron count error of 6.56%. We also demonstrated that the proposed MIMO U-Net can perform equally well when compared to a U-Net equipped with memory units on a publicly available dataset.

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.

REFERENCES

[1] Štajduhar, A., Lepage, C., Judaš, M., Lončarić, S., and Evans, A. C., "3d localization of neurons in bright-field histological images," in [*2018 International Symposium ELMAR*], 75–78, IEEE (2018).

[2] Oberlaender, M., Dercksen, V. J., Egger, R., Gensel, M., Sakmann, B., and Hege, H.-C., "Automated three-dimensional detection and counting of neuron somata," *Journal of neuroscience methods* **180**(1), 147–160 (2009).

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

[4] Mathew, B., Schmitz, A., Muñoz-Descalzo, S., Ansari, N., Pampaloni, F., Stelzer, E. H., and Fischer, S. C., "Robust and automated three-dimensional segmentation of densely packed cell nuclei in different biological specimens with lines-of-sight decomposition," *BMC bioinformatics* **16**(1), 1–14 (2015).

[5] Mazzamuto, G., Costantini, I., Neri, M., Roffilli, M., Silvestri, L., and Pavone, F. S., "Automatic segmentation of neurons in 3d samples of human brain cortex," in [*International Conference on the Applications of Evolutionary Computation*], 78–85, Springer (2018).

[6] Li, R., Zhu, M., Li, J., Bienkowski, M. S., Foster, N. N., Xu, H., Ard, T., Bowman, I., Zhou, C., Veldman, M. B., et al., "Precise segmentation of densely interweaving neuron clusters using g-cut," *Nature communications* **10**(1), 1–12 (2019).

[7] Wang, H., Zhang, D., Song, Y., Liu, S., Wang, Y., Feng, D., Peng, H., and Cai, W., "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)*], 228–231, IEEE (2019).

[8] LaTorre, A., Alonso-Nanclares, L., Peña, J. M., and DeFelipe, J., "3d segmentation of neuronal nuclei and cell-type identification using multi-channel information," *Expert Systems with Applications* **183**, 115443 (2021).

[9] Štajduhar, A., Džaja, D., Judaš, M., and Lončarič, S., "Automatic detection of neurons in neun-stained histological images of human brain," *Physica A: Statistical Mechanics and its Applications* **519**, 237–246 (2019).

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

[11] Dave, P., Goldgof, D., Hall, L. O., Kolinko, Y., Allen, K., Alahmari, S., and Mouton, P. R., "A disector-based framework for the automatic optical fractionator," *Journal of Chemical Neuroanatomy* , 102134 (2022).

[12] Christiansen, E. M., Yang, S. J., Ando, D. M., Javaherian, A., Skibinski, G., Lipnick, S., Mount, E., O'Neil, A., Shah, K., Lee, A. K., et al., "In silico labeling: predicting fluorescent labels in unlabeled images," *Cell* **173**(3), 792–803 (2018).

[13] Arbelle, A. and Raviv, T. R., "Microscopy cell segmentation via convolutional lstm networks," in [*2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*], 1008–1012, IEEE (2019).

[14] Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., and Van Der Maaten, L., "Exploring the limits of weakly supervised pretraining," in [*Proceedings of the European conference on computer vision (ECCV)*], 181–196 (2018).

[15] Nafchi, H. Z., Shahkolaei, A., Hedjam, R., and Cheriet, M., "Corrc2g: Color to gray conversion by correlation," *IEEE Signal Processing Letters* **24**(11), 1651–1655 (2017).

[16] Maška, M., Ulman, V., Svoboda, D., Matula, P., Matula, P., Ederra, C., Urbiola, A., España, T., Venkatesan, S., Balak, D. M., et al., "A benchmark for comparison of cell tracking algorithms," *Bioinformatics* **30**(11), 1609–1617 (2014).

[17] Ulman, V., Maška, M., Magnusson, K. E., Ronneberger, O., Haubold, C., Harder, N., Matula, P., Matula, P., Svoboda, D., Radojevic, M., et al., "An objective comparison of cell-tracking algorithms," *Nature methods* **14**(12), 1141–1152 (2017).

[18] Kingma, D. P. and Ba, J., "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980* (2014).

[19] Delgado, P., Sanchez, K., Anderson, A., Patel, R., Alahmari, S., Goldgof, D., Hall, L., and Mouton, P., "Comparison of manual, semi-automatic and fully automatic counts of immunostained neurons in mouse brains," Soc. For Neurosciences (November 8-11, 2021). *in press.*

[20] Patel, R., Alahmari, S., Goldgof, D., Phoulady, H., Dave, P., Hall, L., and Mouton, P., "Stereological analysis of neurodegeneration and neuroinflammation in tg4510 mice using manual and automatic stereology," *Society for Neurosciences* **558** (2019).