CONSISTENT RECURRENT NEURAL NETWORKS FOR 3D NEURON SEGMENTATION.

Felix Gonda, Donglai Wei, Hanspeter Pfister

Harvard University
Cambridge, MA

ABSTRACT

We present a recurrent network for 3D reconstruction of neurons that sequentially generates binary masks for every object in an image with spatio-temporal consistency. Our network models consistency in two parts: (i) local, which allows exploring non-occluding and temporally-adjacent object relationships with bi-directional recurrence. (ii) non-local, which allows exploring long-range object relationships in the temporal domain with skip connections. Our proposed network is end-to-end trainable from an input image to a sequence of object masks, and, compared to methods relying on object boundaries, its output does not require post-processing. We evaluate our method on three benchmarks for neuron segmentation and achieved state-of-the-art performance on the SNEMI3D [1] challenge.

Index Terms— Recurrent Neural Network, Neuron Segmentation, Instance Segmentation, Object Consistency.

1. INTRODUCTION

The field of connectomics aims to reconstruct the brain's wiring diagram by mapping the neural connections at the level of individual synapses. A reconstruction of neurons' anatomical structures and the synaptic connectivity between them can help neuroscientists better understand the structure and function of the brain [2]. Recent advances in Electron Microscopy (EM) technology make it possible to generate terabytes of brain images at the nanometer scale on an hourly basis [3]. Thus, efficient and accurate neuron segmentation methods are required to process these images.

Previous approaches [4, 5] addressed neuron segmentation in multiple steps. First, a convolutional neural network (CNN) is applied to predict neurons' instance boundaries or affinities. A watershed transform [6] is then used to generate an initial 3D over-segmentation, where segments are further merged based on hand-crafted or learned features. However, the CNNs are applied independently on each local subvolume without any shape information about neighboring regions. Thus, when image artifacts or unexpected appearance occurs in the input volume, the CNNs make wrong predictions, leading to merging and splitting errors in the final segmentation.

More recent approaches [7, 8] treat neuron segmentation as video object tracking along the z-axes of a 3D image volume. These approaches alleviate the appearance problem by learning spatial features through recurrent neural network models. These models segment one object and generate the object mask for the input sub-volume with additional input features from the previous inference step. However, these methods are computationally expensive. For instance, on a 9x9x20nm image volume of the zebra finch songbird dataset, one inference step of the flood-filling networks [7] costs 41.05 EFLOPS with a wall time of 3.15 hrs.

In terms of spatial consistency, the affinity-based methods [4, 5] learn them all-in-one, while the tracking-based methods [7, 8] learn them one-by-one without exploiting the pairwise non-occluding relationships. In terms of temporal consistency during inference, the tracking-based methods run the inference only in forward and backward directions, without long-range consistency. As such, image slices with severe artifacts can lead to broken segmentation. Our approach is inspired by the idea of spatio-temporal recurrence previously demonstrated for video object segmentation [9], where objects are segmented by propagating masks from a reference frame. We argue that exploiting the reference frame's semantic meaning has benefits in learning long-range object relationships in sequences. As such, we aim to extend spatial and temporal recurrence to explore local and non-local relationships between objects in the volume to connect broken segments and form more complete object masks.

Therefore, we propose a recurrent model that performs a sequential analysis of the input image volume to deal with complex object distributions and generate consistent predictions. We create a new recurrent module based on Convolutional LSTM [10], as a building block of our network, to model local and non-local object consistency. Given a single label map, our model can segment multiple objects without post-processing. We require no intermediate representation for our model; as such, its training is performed in an end-to-end fashion. Accuracy results on three connectomics datasets show that our method performs similar to state-of-the-art methods and surpasses human accuracy for 3D neuron segmentation on the SNEMI3D [1] dataset. Our results also demonstrate that our method consistently produces object masks that are robust to image artifacts.

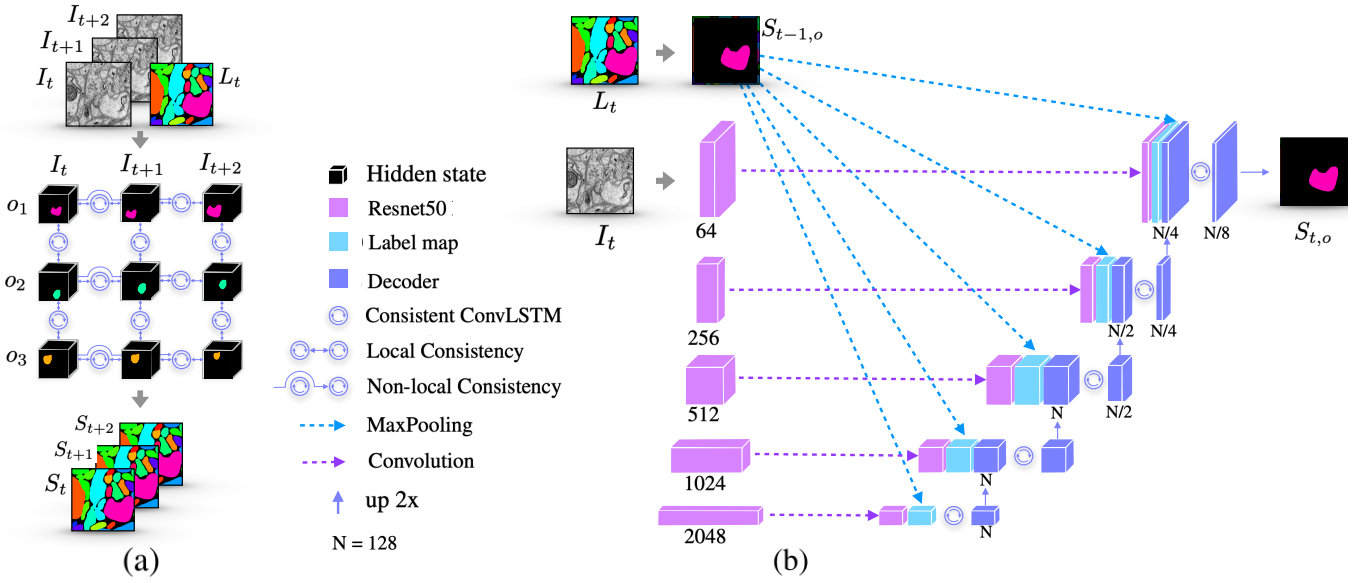

Fig. 1. Our proposed architecture for multi-neuron segmentation. (a) A network of our Consistent ConvLSTM extract neurons from a sequence I based on a single label map L_t. (b) A single forward pass of our model predicting the first object $S_{t,o}$ of a frame given the mask of the object from the previous frame $S_{t-1,o}$. When t is the reference frame, $S_{t-1,o} = L_{t,o}$

2. METHODOLOGY

2.1. Consistent Recurrent Neural Network

We propose a recurrent network, depicted in Fig 1, based on an encoder-decoder architecture to solve the task of neuron segmentation. The network incorporates spatio-temporal recurrence with our Consistent ConvLSTM (CConvLSTM) module defined in Section 2.2. Our recurrence is configured in the spatial domain (rows) to represent the object instances in a frame and in the temporal domain (columns) to represent frames. The model's input consists of 3D patches transformed into sequences of images along the z-direction of the input volume and a single label map for each sequence. For refinement, each sequence is accompanied by a channel carrying initial object mask estimation. The output of our model is a set of predicted masks $S = \{S_{t,o}, S_{t,o+1}, ..., S_{t,o+M}\}$, where $S_{t,o}$ is the predicted mask of object o at frame t and M is the number of objects.

2.2. Spatio-Temporal Consistency Module

To model object consistency, we introduce a new recurrent module, CConvLSTM shown in Fig. 2, as a building block of our network. This module models local object consistency by combining two ConvLSTM layers that process input in the forward and backward directions. The output of the two layers is convolved with the reference hidden state to model non-local object consistency. The bi-directional flow analysis in the temporal domain has been shown in [11] to yield improved predictive performance.

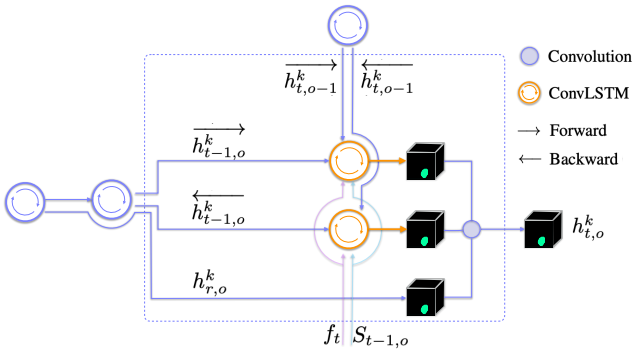

Fig. 2. A CConvLSTM module models local object consistency with two ConvLSTM layers to process input in the forward and backward directions. Non-local object consistency is modeled with a skip connection that integrates the reference hidden state $h^k_{r,o}$

Therefore, given a sequence and a set of objects propagated from a reference frame r, the output $h^k_{t,o}$ of the k^{th} CConvLSTM layer for object o at frame t is computed as:

$$h_{input} = [B_2(h^{k-1}_{t,o})|f^{k'}_t|S_{t-1,o}] \quad (1)$$

$$\overrightarrow{h_{state}} = [\overrightarrow{h^k_{t,o-1}}|\overrightarrow{h^k_{t-1,o}}] \quad (2)$$

$$\overleftarrow{h_{state}} = [\overleftarrow{h^k_{t,o-1}}|\overleftarrow{h^k_{t-1,o}}] \quad (3)$$

$$h^k_{t,o} = CConvLSTM(h_{input}, \overrightarrow{h_{state}}, \overleftarrow{h_{state}}, h^k_{r,o}) \quad (4)$$

where, B2 is the bilinear upsampling of the output of the pre-

vious CConvLSTM layer by a factor of 2. f_t^k is the features from the encoder at frame t and $f_{t,k}^{k'}$ is the projection of f_t^k to lower dimension via a convolution layer. $S_{t-1,o}$ is the predicted segmentation mask of the object from the previous frame. $\overrightarrow{h_{t,o-1}^k}$ and $\overleftarrow{h_{t,o-1}^k}$ are the forward and backward components of the spatial hidden state representation of the previous object. $\overrightarrow{h_{t-1,o}^k}$ and $\overleftarrow{h_{t-1,o}^k}$ are the forward and backward components of the temporal hidden state of the object from the previous frame. For the first hidden state $h_{t,o}^0$ of the object, we use the segmentation mask from the reference $S_{r,o}$ frame and the zero matrix for the spatial hidden state. We assess the importance of the local and non-local object consistency in Section 3.2.

2.3. Encoding Path

The encoder, illustrated in pink in Fig. 1, is a Resnet50 [12] that is truncated at the last convolution layer. The encoder learns to extract features, $f = \{f_t, f_{t+1}, f_{t+2}, f_{t+3}, f_{t+4}\}$, from an RGB image $x \in R^{h \times w \times 3}$ corresponding to the output of Resnet blocks. f_t corresponds to the output of the deepest block, and f_{t+4} corresponds to the output of the block whose input is the image. A convolution operation is used to extract the features from each block.

2.4. Decoding Path

The decoder, shown in dark blue in Fig. 1 (b), is a hierarchical recurrent architecture of CConvLSTMs leveraging the different resolutions of the input features f and label map L_t. The label map is extracted with a series of down-sampling operations corresponding to the resolution of features f as shown in blue in Fig. 1 (b). The output of each CConvLSTM is subsequently merged with corresponding encoder features and object masks, which allows the decoder to reuse low-level features and refine the final segmentation. The decoder applies equation 4 in chains for the number of CConvLSTMs. The decoder's output is a set of M predictions per image. M is the number of objects propagated and is always constant per sequence to ensure an object mask will be empty if it disappears in subsequent frames. The constant M also ensures the predicted mask in the temporal domain is consistent with the spatial recurrence.

2.5. Implementation Details

Since the number of propagated objects is always equal to the number of predicted objects, we estimate the parameters of our model by optimizing an objective function based on the Hungarian algorithm [13] using soft intersection over union. Thus, given a sequence of m predicted and g ground truth masks of length N, the loss can be expressed as:

$$sIoU(m, g) = 1.0 - \frac{\sum_{i=1}^{N} m_i g_i}{(\sum_{i=1}^{N} m_i + g_i - m_i g_i)}$$

The network is trained using the Adam optimizer with a learning rate of 10^{-6} and a batch size of 1 over 40 epochs. For the early ten epochs, the ground-truth mask of objects from the previous frame is included as an additional input channel to our CConvLSTM. For the remaining 30 epochs, the object's inferred mask is used to fine-tune the model, thus allowing the model to learn to fix errors that may occur at inference time. The training was carried out on a single NVIDIA Titan X GPU with 12GB RAM for 24 hours.

3. EXPERIMENTS

3.1. Setup

We evaluate our method's efficacy with experiments on three EM datasets from different species, as described in Table 1. The (x, y, z) dimensions of each datasets in voxels is: $(1024 \times 1024 \times 100)$, $(500 \times 500 \times 500)$, and $(1024 \times 1024 \times 105)$ for SNEMI [1], FIBSEM [14], and FIBER respectively. The FIBER dataset is in-house generated, the FIBSEM is public, and the SNEMI is a benchmark for the SNEMI challenge.

Name	Species (region)	Volume
SNEMI [15]	Mouse (Cortex)	$6 \times 6 \times 3\,\mu\text{m}^3$
FIBSEM [14]	Fruit Fly (Medulla)	$5 \times 5 \times 5\,\mu\text{m}^3$
FIBER (in-house)	Mouse (Cerebellum)	$8 \times 8 \times 3\,\mu\text{m}^3$

Table 1. The list of datasets used to evaluate our system. Each consists of a training and testing volume. The training volume is split 80% for training and 20% for validation.

Our baseline model implements spatio-temporal (ST) recurrence. We consider three options to analyze the importance of consistency: (i) local consistency model STL (non-local consistency not used), (ii) non-local consistency model STN (local consistency not used), and (iii) a combined model STC (both local and non-local consistency used). We compare our best model against the SNEMI challenge leaderboard and the affinity-based model, waterz [4]. We use the author's publicly available waterz method implementation and use the original paper's suggested hyper-parameters.

During inference, we process the input 3D volume patches in an overlapping manner, starting with a labeled frame. The segmentation of the first frame is initialized with its corresponding ground-truth. A watershed transform output is used to initialize the first frame's segmentation for volumes where no ground-truth exists.

We analyze our results using the Adaptive Rand Index (ARI) [16] used by the SNEMI challenge [1] to be consistent. The ARI measures the similarity between two data clusters. The error is defined as one minus the maximal F-score of the Rand index. A lower ARI score corresponds to better segmentation quality.

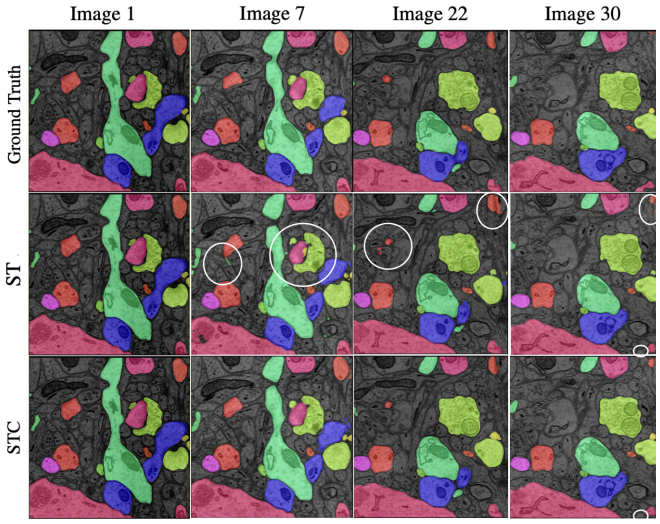

Fig. 3. We extract 15 neurons from a sequence of 30 images from the SNEMI dataset using the ST and our STC models. In comparison to the ground truth, errors are circled in white.

3.2. Results

In Fig. 3, we show the segmentation of 15 objects by the ST and STC models in a validation sequence compared to ground-truth. As shown by the white circles, propagation errors occur when consistency is not used. These observations are further confirmed with the quantitative evaluations, shown in Table 2. The non-local consistency model (STN) is faster and connects distant segments better, and when combined with local consistency, produces the lowest ARI value as demonstrated by the STC model shown in Table 2.

Model	Accuracy (ARI)	Inference Time (seconds)
ST [9]	0.13	520.0
STL	0.082	640.0
STN	0.045	605.0
STC	0.035	660.0

Table 2. An assessment of local (L) and non-local (N) consistency compared to baseline ST [9] model in terms of ARI (lower is better) and inference time in seconds.

For a fair comparison against state-of-the-art methods, we start from the same watershed output as the waterz[4] method. In Fig. 4, we highlight in red circles errors in the waterz method that are addressed by our spatio-temporal consistency. Although our method is robust to segmentation breakage, some artifacts may not be fully corrected as shown in black if the watershed is bad. In this case, fine-tuning the watershed before applying our model is an alternate solution. In table 3, we compare our best model STC against the SNEMI3D [1] challenge leaderboard. Our method achieved 0.035 ARI score, surpassing the human accuracy, and is currently ranked third after PNI [5] and FNN [7]. Our method also consistently outperformed the waterz[4] method on all three datasets. We attribute our method's strength to its ability to learn long-range object relationships, critical for connectomics data. This is demonstrated, in Figure 4, with the presence of segment splits in waterz and their absense in our method. In terms of inference time on the SNEMI dataset, our method requires 660 seconds compared to 61 minutes for the FFN [7] model.

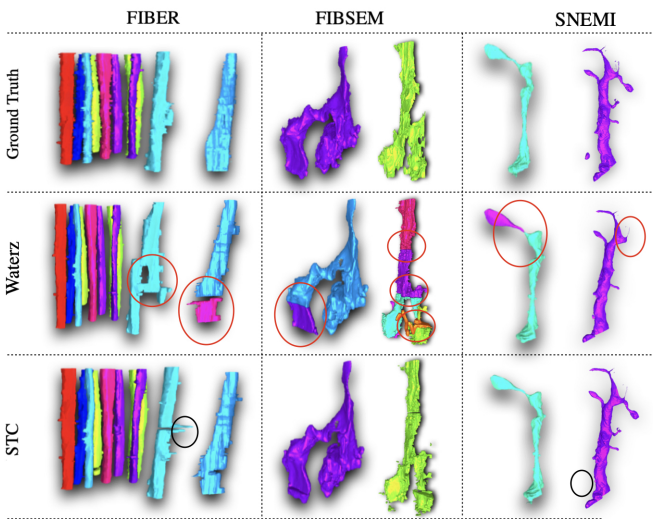

Fig. 4. Qualitative comparison of our STC to the waterz [4] method. Errors shown in red and black circles.

Method/Dataset	SNEMI3D	FIBSEM	FIBER
PNI [5]	**0.024**	N/A	N/A
FFN [7]	**0.029**	N/A	N/A
STC (ours)	**0.035**	**0.106**	**0.091**
(human values)	0.059	N/A	N/A
waterz [4]	0.072	0.163	0.210

Table 3. Our STC achieved third place ranking on SNEMI3D [1] leaderboard and outperform waterz [4] on all three datasets.

4. CONCLUSION

The proposed recurrent model learns to deal with complex object distribution across long sequences and produces segments that are consistent with each other. By training our model end-to-end, we eliminated intermediate representations that could potential introduce errors such as in the affinity-based methods. Our model is also suited for interactive segmentation where object propagation is driven by the user. Therefore, we plan to incorporate our model in the proofreading of neurons to help correct automatic segmentation errors.

5. COMPLIANCE WITH ETHICAL STANDARDS

This is a numerical simulation study for which no ethical approval was required.

6. ACKNOWLEDGMENTS

This work was partially supported by NSF grant IIS-1607800.

7. REFERENCES

[1] Ignacio Arganda-Carreras, Srinivas C. Turaga, Daniel R. Berger, Dan Cireşan, Alessandro Giusti, Luca M. Gambardella, Jürgen Schmidhuber, Dmitry Laptev, Sarvesh Dwivedi, Joachim M. Buhmann, Ting Liu, Mojtaba Seyedhosseini, Tolga Tasdizen, Lee Kamentsky, Radim Burget, Vaclav Uher, Xiao Tan, Changming Sun, Tuan D. Pham, Erhan Bas, Mustafa G. Uzunbas, Albert Cardona, Johannes Schindelin, and H. Sebastian Seung, "Crowdsourcing the creation of image segmentation algorithms for connectomics," *Frontiers in Neuroanatomy*, vol. 9, pp. 142, 2015.

[2] Joshua L Morgan and Jeff W Lichtman, "Why not connectomics?," *Nature methods*, vol. 10, no. 6, pp. 494, 2013.

[3] Richard Schalek, Dongil Lee, Narayanan Kasthuri, Adi Suissa-Peleg, Thouis R. Jones, Verena Kaynig, Daniel Haehn, Hanspeter Pfister, David Cox, and Jeffery W. Lichtman, "Imaging a 1 mm3 volume of rat cortex using a multibeam sem," in *Microscopy and Microanalysis*. 26 July 2016, vol. 22, pp. 582–583, Cambridge Univ Press.

[4] Jan Funke, Fabian Tschopp, William Grisaitis, Arlo Sheridan, Chandan Singh, Stephan Saalfeld, and Srinivas C. Turaga, "A deep structured learning approach towards automating connectome reconstruction from 3d electron micrographs," *CoRR*, vol. abs/1709.02974, 2017.

[5] Kisuk Lee, Jonathan Zung, Peter Li, Viren Jain, and H. Sebastian Seung, "Superhuman accuracy on the SNEMI3D connectomics challenge," *CoRR*, vol. abs/1706.00120, 2017.

[6] Aleksandar Zlateski and H. Sebastian Seung, "Image segmentation by size-dependent single linkage clustering of a watershed basin graph," *CoRR*, vol. abs/1505.00249, 2015.

[7] Michal Januszewski, Jeremy Maitin-Shepard, Peter Li, Jörgen Kornfeld, Winfried Denk, and Viren Jain, "Flood-filling networks," *CoRR*, vol. abs/1611.00421, 2016.

[8] Yaron Meirovitch, Alexander Matveev, Hayk Saribekyan, David Budden, David Rolnick, Gergely Ódor, Seymour Knowles-Barley, Raymond Thouis Jones, Hanspeter Pfister, William Jeff Lichtman, and Nir Shavit, "A multi-pass approach to large-scale connectomics," *arXiv: Quantitative Methods*, 2016.

[9] C. Ventura, M. Bellver, A. Girbau, A. Salvador, F. Marques, and X. Giro i Nieto, "Rvos: End-to-end recurrent network for video object segmentation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA, USA, jun 2019, pp. 5272–5281, IEEE Computer Society.

[10] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," *CoRR*, vol. abs/1506.04214, 2015.

[11] Zhiyong Cui, Ruimin Ke, and Yinhai Wang, "Deep bidirectional and unidirectional LSTM recurrent neural network for network-wide traffic speed prediction," *CoRR*, vol. abs/1801.02143, 2018.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.

[13] H. W. Kuhn and Bryn Yaw, "The hungarian method for the assignment problem," *Naval Res. Logist. Quart*, pp. 83–97, 1955.

[14] Shin-ya Takemura, C Shan Xu, Zhiyuan Lu, Patricia K Rivlin, Toufiq Parag, Donald J Olbris, Stephen Plaza, Ting Zhao, William T Katz, Lowell Umayam, et al., "Synaptic circuits and their variations within different columns in the visual system of drosophila," *PNAS*, 2015.

[15] Narayanan Kasthuri, Kenneth Jeffrey Hayworth, Daniel Raimund Berger, Richard Lee Schalek, José Angel Conchello, Seymour Knowles-Barley, Dongil Lee, Amelio Vázquez-Reina, Verena Kaynig, Thouis Raymond Jones, et al., "Saturated reconstruction of a volume of neocortex," *Cell*, vol. 162, no. 3, pp. 648–661, 2015.

[16] R. Unnikrishnan, C. Pantofaru, and M. Hebert, "Toward objective evaluation of image segmentation algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 929–944, June 2007.