










# Robust and Markerfree *in vitro* Axon Segmentation with CNNs

Philipp Grüning<sup>1</sup> , Alex Palumbo<sup>2,3,4</sup> , Svenja Kim Landt<sup>2,3</sup> ,  
Lara Heckmann<sup>2,3</sup> , Leslie Brackhagen<sup>1</sup> , Marietta Zille<sup>2,3,4</sup> ,  
and Amir Madany Mamlouk<sup>1</sup> 

<sup>1</sup> Institute for Neuro- and Bioinformatics, University of Lübeck, Lübeck, Germany  
{gruening,madany}@inb.uni-luebeck.de

<sup>2</sup> Fraunhofer Research and Development Center for Marine and Cellular  
Biotechnology, Lübeck, Germany

<sup>3</sup> Institute for Medical and Marine Biotechnology, University of Lübeck, Lübeck,  
Germany

<sup>4</sup> Institute for Experimental and Clinical Pharmacology and Toxicology,  
University of Lübeck, Lübeck, Germany

**Abstract.** The automated *in vitro* segmentation of axonal phase-contrast images to allow axonal tracing over time is highly desirable to understand axonal biology in the context of health and disease. While deep learning has become a powerful tool in biomedical image analysis for semantic segmentation tasks, segmentation performance has been limited so far since axons are long and thin objects that are sensitive to under- and/or over-segmentation. We here propose the use of an ensemble-based convolutional neural network (CNN) framework for the segmentation of axons on phase-contrast microscopic images. The mean ResNet-50 ensemble performed better than the max u-net ensemble on the axon segmentation task. We estimated an upper limit for the expected improvement using an oracle-machine. Additionally, we introduced a soft version of the Dice coefficient that describes the visually perceived quality of axon segmentation better than the standard Dice. Importantly, the mean ResNet-50 ensemble reached the performance level of human experts. Taken together, we developed a CNN to robustly segment axons in phase-contrast microscopy that will foster further investigations of axonal biology in health and disease.

**Keywords:** Axon segmentation · Microscopy · Ensembles · ResNet-50

## 1 Introduction

Axons are wire-like extensions from neuronal cell bodies that ensure the communication to neighboring neurons by building connections among them. Axonal morphology is highly complex, with varying lengths, diameters, and degrees of arborization [3] and studying the role of axons in health and disease is a major emphasis of current research [10].

In cell culture, individual axonal structures can be followed over longer periods of time by time-lapse microscopy. The respective data analysis, however, requires dedicated software tools that allow for the precise identification of axonal structures. At the same time, these software tools need to cope with the large amount of data available from imaging where manual inspection is time consuming, prone to error, and impractical [1, 12].

Over the past two decades, many software packages such as NeuronMetrics, NeuriteIQ, NeuriteTracer, and NeurphologyJ have been developed to trace axons [7, 12]. All of these tools are able to trace axonal structures only semi-automatically and require high-contrast images that are only available in fluorescence and not in phase-contrast microscopy. Apart from bleaching issues, fluorescence imaging requires either fixation of the cells, which limits the observation to a single time point, or genetic modulation, which is less efficient in primary cells and may alter the behavior of the cells. Another tool, NeuronGrowth, is able to analyze live-cell imaging recordings, but also needs user intervention to select the starting point of the axonal structures to be traced [5]. Thus, automated software that allows for axonal tracing over time, based on phase-contrast images – as it is well-established for *in vitro* cell tracking [16] – is highly desirable and will greatly enhance our understanding of axonal function and morphology in health and disease.

Many approaches to automatically segment axons are based on traditional image processing algorithms, including global thresholding, Laplacian or Gaussian filters, and morphological operations [13]. These approaches come with a number of drawbacks: i) They are static and do not react robustly to changes in data collection or the hardware used, ii) most of these procedures are adapted to a particular application scenario and it is unlikely that they generalize well across a wide range of experimental setups and questions, and iii) they are therefore semi-autonomous, i.e., user interaction is required before the data can be collected and automatically evaluated. As axons display morphological variability, the complete segmentation of such an object is a highly demanding task.

In recent years, deep learning has expanded horizons in the field of image processing, ranging from image classification [8] to more intricate tasks such as detection or semantic segmentation with fully convolutional networks (FCNs) [11]. Especially in biomedical image analysis, a very common FCN architecture for segmentation tasks is the u-net [17]. In many cases, the segmentation performance of a network can be further increased using transfer learning [9], i.e., employing deeper architectures such as ResNets [6] that were previously trained on a demanding dataset, e.g., Imagenet [4].

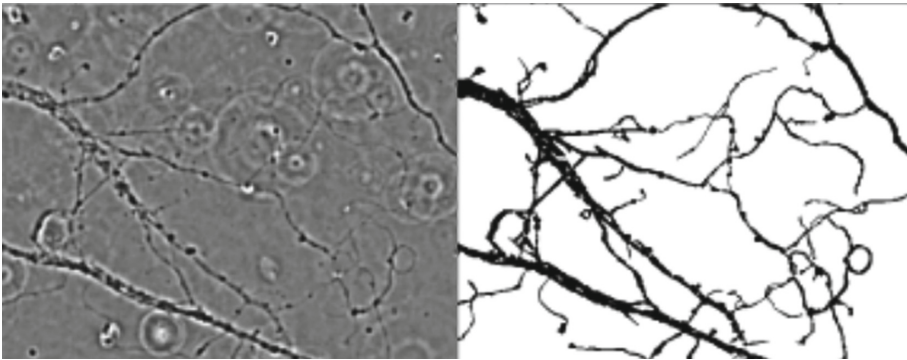
There are few studies that have applied CNNs on axon segmentation in 2D [14, 15, 18] and 3D [20]. However, to our knowledge, there are no works on 2D phase-contrast microscopy images that enable the automated segmentation of axonal morphology over time.

In this work, we used CNNs to robustly and reliably segment axons on marker-free phase-contrast microscopic images in an automated manner. We employed an ensemble approach to improve the quality of the output and estimated an

upper limit for the expected improvement using an oracle-machine. We introduced a soft version of the Dice coefficient that describes the visually perceived quality of axon segmentation better than the standard Dice. Finally, we demonstrate that our best model already reaches the performance level of human experts.

## 2 Data and Methods

**Data.** We used microfluidic devices to separate neuronal cell bodies from their axons [2]. We isolated murine primary cortical neurons from embryonic day 14.5 from Crl:CD1 (ICR) Swiss outbred mice (Charles River) as previously described [21] (under the prospective contingent animal license number 2017-07-06\_Zille approved by the Schleswig-Holstein Ministry for Energy Transition, Agriculture, Environment, Nature and Digitalization). We seeded the cells to one compartment of the device, which extended their axons through the microgrooves to the other compartment due to the volume difference of the two compartments. We captured grayscale images of the axonal compartment using an Olympus IX2 inverted microscope from which 42 images were manually labeled using GIMP v.2.10.14 (GNU Image Manipulation Program, RRID:SCR\_003182). Each image had a size of  $1200 \times 1000$  pixels on average. Figure 1 shows an example of the data.



**Fig. 1.** Original image and binary label image: The left picture shows a pre-processed section of the original data, i.e., microscopic images of axons. The corresponding (manually drawn) binary mask can be seen on the right image, which denotes all pixels that are part of an axon in the left image.

**Network Training and Ensembles.** We compared two architectures: a standard u-net [17] and a u-net with a ResNet-50 encoder [6]. For each architecture, we trained 8 networks on 10 *splits*. For each split, we separated the dataset into 31 training images and tested on 11 images. Both architectures were trained for

90 epochs with stochastic gradient descent, a momentum of 0.9 and a learning rate of 0.1. Every 30 epochs, we decreased the learning rate by a factor of 10. We used a batch size of 4. For data augmentation, images were cropped randomly with an input size of  $512 \times 512$  pixels. Different input sizes did not alter the performance and the size of 512 pixels exceeds the receptive field of both networks.

By training 8 networks per split, we generated 8 output maps for each test image. We compared the pixel-wise mean with the pixel-wise maximum and the best single model of each split (*mean-ensemble*, *max-ensemble*, *best model*).

**Oracle Machine.** To investigate the impact of different ensemble strategies, we used an oracle-machine. The max-ensemble achieved a much better recall than the mean-ensemble (0.900 versus 0.799 for ResNet-50). Therefore, we defined a max-mean-oracle for the two critical cases when both, max- and mean-ensemble disagreed in their decision: If the max-ensemble recognized an axon but the mean-ensemble did not (false negatives for the mean-ensemble) and - vice versa - if the max-ensemble was wrong (false positives for the mean-ensemble). As the oracle-machine can perfectly distinguish both cases, we used this oracle to estimate an upper limit on how good the performance would be when combining the information from both ensemble strategies.

**$\varepsilon$ -Dice Score.** We based our evaluation on the standard Dice score. But even if the prediction-label pairs looked reasonable on visual inspection, the Dice score can be low. To test whether areas were just missed or simply not detected at all, we used a soft version of the Dice score, called  $\varepsilon$ -Dice: If a ground truth pixel was within the proximity of a false positive prediction (i.e., in a neighborhood of  $\varepsilon$  pixels), we defined this false positive as an over-segmented true positive. Thus, axon predictions that were slightly thicker than the ground truth mask were not counted as errors. False negatives were defined as under-segmented true positives, if there was another true positive in the given neighborhood.

Note that the  $\varepsilon$ -Dice requires explicit knowledge of the ground truth and thus did **not** improve the accuracy of the segmentation in any way. Rather, we used this measure to estimate how much of the error occurred in the immediate vicinity of the axons or whether there were completely undetected axon segments.

**Comparison to Human Performance.** To further test our assumption that a perfect Dice score is almost impossible to accomplish on this dataset, we compared it to the human performance level on this task using 7 images labeled by three experts, which we compared to each other and to our best model.

### 3 Results

**The Mean ResNet-50 Ensemble Outperformed the U-Net Ensemble and All Single Networks.** To identify the best performing CNN, we compared the mean- and max-ensembles as well as the best single models of each split for both u-net and ResNet-50 (see Table 1). We observed that the ResNet-50 was

superior to the u-net with the mean ResNet-50 giving the best results (Dice = 0.827). To estimate the influence of the pixels for which max- and mean-ensemble would vote differently, we defined an oracle-ensemble, which would always make the right decision in these cases. The oracle improved the Dice score for both u-net and ResNet-50 by about 20 %.

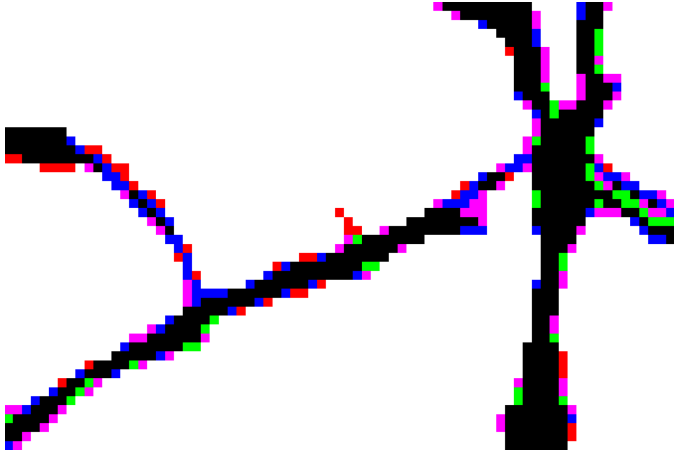
**Table 1.** Dice score for 10 train and test splits with the u-net and ResNet-50. *bestmodel* was the best of the 8 single networks. *max* always used the highest output from the ensemble. *mean* was rated based on the average ensemble rating. With different ratings of *max* and *mean*, *oracle* always made the correct decision. The results are shown for u-net and ResNet-50 using normal Dice score ( $\varepsilon = 0$ ) and a soft Dice ( $\varepsilon = 1$ ).

Method	u-net		ResNet-50 ( $\varepsilon = 0$ )		ResNet-50 ( $\varepsilon = 1$ )	
	mean	std	mean	std	mean	std
bestmodel	0.757	0.042	0.815	0.023	0.939	0.018
max	<b>0.784</b>	<b>0.034</b>	0.805	0.029	0.927	0.028
mean	0.754	0.046	<b>0.827</b>	<b>0.021</b>	<b>0.942</b>	<b>0.016</b>
oracle	0.852	0.028	0.887	0.014	0.965	0.011

**Segmentation Errors Occurred on the Object Border.** Comparing the original mask (ground truth) and the resulting masks from the different approaches, the potential errors occurred at the edges of the object. Upon closer examination, we revealed that the critical pixels were located more or less randomly at the object edges (Fig. 2) and it was hardly ever the case that a whole section of an axon was not segmented (Fig. 3). We did not find any further scheme that was able to distinguish over- or under-segmentation here.

**The  $\varepsilon$ -Dice Described Best the Visually Perceived Quality of Axon Segmentation.** To test whether the observed segmentation error can be attributed to cumulative individual errors, we used the  $\varepsilon$ -Dice score that also includes the surrounding pixels in the evaluation. We observed that the  $\varepsilon$ -Dice exceeded 90 % for all examined approaches, with the ResNet-50 mean-ensemble achieving the best result with 94 % (Table 1). Also noteworthy is the reduction in the distance between our ensemble approach and the oracle to only 2 %, indicating that many of the critical pixels were located close to uncritical axon structures.

**The Recall-Precision Trade-Off Can Be Altered by Linear Classification.** We observed that the max-ensemble achieved a better recall than the mean-ensemble (0.900 vs. 0.799). Therefore, we investigated if combining the max-recall with the mean resulted in a better segmentation. When both approaches made the same decision, the performance did not improve. However, two cases are critical (Fig. 4): If the max-ensemble recognized an axon but the mean-ensemble did not (mean false negatives, case a) and - vice versa - if the max-ensemble is



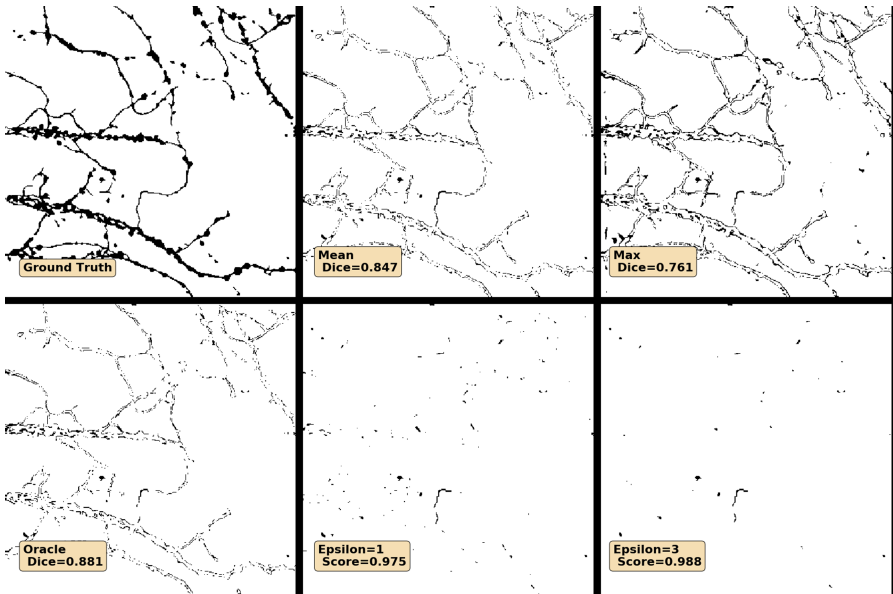
**Fig. 2.** Comparison of the mean- and max-ensemble and the oracle (best viewed in color). Black pixels show true positives correctly segmented by all approaches. Green and red pixels show false positives and false negatives of all approaches, respectively. The critical pixels that can further improve the result are shown in blue. Here, the max-ensemble yielded a true positive prediction, while the mean-ensemble predicted a false negative. However, the output of the max-ensemble increased the number of false positives (pink pixels). (Color figure online)

wrong (max false positives, case b). The distribution for case a) indicated that for many pixels, the max score was close to 1.0, the mean score was close to 0.5 but did not exceed it. In case b), however, the mean score was rather small ( $< 0.2$ ) and the max-score was only slightly above 0.5.

Thus, we defined a 2-dimensional linear classifier that re-determined the output of the ensemble for those relevant pixels (Fig. 4). We evaluated the results for three linear classifiers, where each separating line was orthogonal to the manually determined line spanned between  $\mathbf{p}_0 = (0.05, 0.5)^T$  and  $\mathbf{p}_1 = (0.5, 1.0)^T$ . The three classifiers had the same normal vector  $\mathbf{n} = (\mathbf{p}_1 - \mathbf{p}_0) / \|\mathbf{p}_1 - \mathbf{p}_0\|^2$ , but differed in their bias value  $b \in \{0.3, 0.5, 0.7\}$ . Note that  $b$  can be seen as the percentage of the distance between  $\mathbf{p}_0$  and  $\mathbf{p}_1$ . The decision is reached as follows:

$$f(\mathbf{x}) = \begin{cases} 1 & \text{if } (\mathbf{x} - \mathbf{p}_0)(\mathbf{n}) \geq b \\ 0 & \text{else} \end{cases} . \quad (1)$$

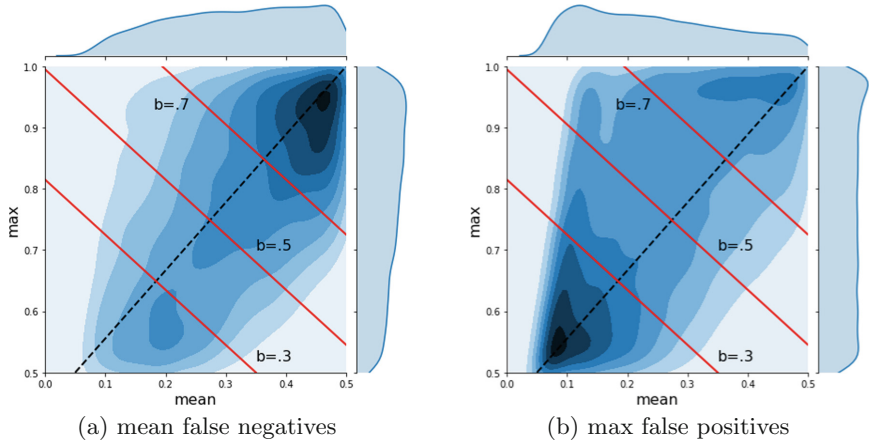
The first two approaches achieved a better recall than the mean-ensemble, but since the precision decreased similarly, the overall Dice-score did not change (Fig. 5). The third approach was almost identical to the mean-ensemble, and again, the quality of the segmentation did not improve.



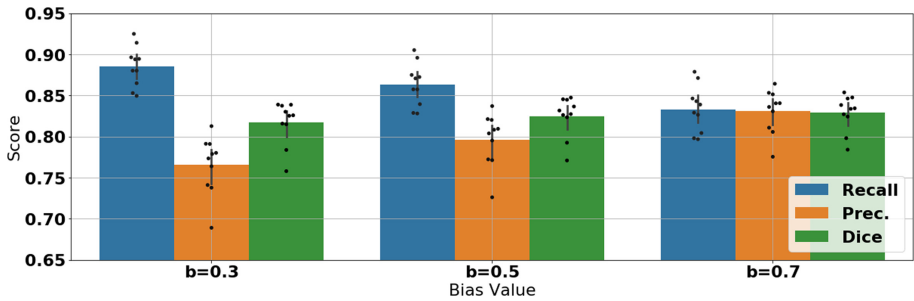
**Fig. 3.** Ground truth and error images of the different approaches and metrics: In all images except the ground truth image, a black pixel indicates a deviation from the label. The two images on the bottom middle and bottom right show the counted error pixels of the mean-ensemble when the  $\epsilon$ -Dice score was used. Although, the max-ensemble increased the recall, a strong over-segmentation decreased the precision. The oracle further detected the few axon pixels that were correctly predicted by the max-ensemble but were neglected by the mean-ensemble. The  $\epsilon$ -Dice images show that the majority of errors was due to over- and under-segmentation and only scarcely, small isolated regions were misclassified.

### The Mean ResNet-50 Ensemble Reached Human Expert Performance.

Finally, in addition to the expert that labeled the entire data set, we asked two more experts to re-label some of the images used here to examine the variance in their ratings (Table 2). Thus, we had the opinions of three experts for evaluation and the test segmentation results of a ResNet-50 ensemble. These experts among themselves hardly achieved a better result than the mean ResNet-50 ensemble. On the contrary, none of the other experts came as close to the masks of the author of the training data (Expert 02) as the CNN ensemble (Dice = 0.793, 0.766, and 0.833 for 01, 03, and mean ResNet-50 vs. 02). This highlights that our approach can sensitively and specifically segment axons on phase-contrast microscopic images at a level similar to manual labeling by experts and is thus suitable for further application.



**Fig. 4.** Distribution of critical pixels on mean- and max-ensembles (darker shades indicating a higher point density). All pixels that were rated differently by the mean- and max-ensemble were considered critical. There were two cases: **(a)** the pixels that the mean-ensemble did not recognize as axons (mean false negatives) and **(b)** those that the max-ensemble incorrectly recognized as axons (max false positive).



**Fig. 5.** Scores for different line values. For three different configurations, a linear classifier for the critical points was chosen and decided after its voting. The critical data was mixed and consequently, either recall or precision improved for each of the settings, but without improving the Dice score significantly.

**Table 2.** Dice score comparison of different human annotators and our best approach (mean ResNet-50). Note that 02 annotated the training data for the network.

Name	01	02	03	mean ResNet-50
01	1.000 ± 0.000	0.793 ± 0.016	0.773 ± 0.035	0.794 ± 0.026
02	0.793 ± 0.016	1.000 ± 0.000	0.766 ± 0.040	0.833 ± 0.033
03	0.773 ± 0.035	0.766 ± 0.040	1.000 ± 0.000	0.750 ± 0.043
mean ResNet-50	0.794 ± 0.026	0.833 ± 0.033	0.75 ± 0.043	1.000 ± 0.000



## 4 Discussion

We here present an ensemble-based CNN framework for the automatic segmentation of axons on phase-contrast microscopic images. We demonstrate that the ResNet-50 is superior to the u-net and that an ensemble can further improve the results. Importantly, our approach reaches the performance level of human experts.

As axons are thin and highly branched objects, segmentation is difficult and thus we needed to use a very deep network (ResNet-50) to reach the best performance. The ResNet-50 outperformed the u-net because i) ImageNet pretraining leads to better features and a better starting point in parameter space, ii.) deeper architectures generalize better, and iii) residual connections enable the learning of identity mappings [6, 11].

Interestingly, even with an ensemble of multiple ResNets, we achieved a Dice score of about only 83%, despite the fact that visually inspected results looked very convincing. While the Dice score is a widely used measure to evaluate segmentation results, here, the cumulative errors in the very close proximity of the axons induced a strong bias. Therefore, we proposed the soft Dice score and were able to demonstrate that 94% of the ground truth within a 1-pixel radius were actually recognized by our ensemble, which we think better reflects the visually perceived performance.

We further demonstrated with the help of an oracle what perfect ensemble recombination can achieve. It would theoretically be possible to reach a Dice score of almost 89% by combining max- and mean-ensemble. However, in practice and as demonstrated by the linear regression model, this seems almost impossible to achieve as we did not identify an approach to combine the knowledge of both ensembles in a usable way. Finally, a comparison with human experts revealed that our ResNet-50 ensemble can very well reach human performance level in the task of axon segmentation.

The dataset used here is relatively small due to the remarkably high labeling cost for these delicate structures, but the size is comparable to similar datasets such as the IOSTAR retina vessel segmentation set [19]. This further strengthens the contribution of our approach as using a network with a performance comparable to an expert can aid in labeling more images.

Taken together, the proposed ensemble CNN allows for the automated axon segmentation at a near-human performance level that makes the high-throughput analysis of the markerfree *in vitro* detection of axonal morphology, growth, and degeneration in health and disease a feasible task.

**Acknowledgments.** This work was supported by a Fraunhofer MEF grant (Project number 600199) to M.Z.

**Contribution of Authors.** P.G. developed the deep learning algorithms and performed the computational experiments and analyzed the data. A.P. performed the *in vitro* experiments. A.P., S.K.L., and L.H. labeled the images. P.G., M.Z., and A.M.M. wrote the manuscript. All authors discussed and commented on the final version of the manuscript.

**Data Availability.** The data and code are available upon reasonable request.

## References

1. Acciai, L., Soda, P., Iannello, G.: Automated neuron tracing methods: an updated account. *Neuroinformatics* **14**(4), 353–367 (2016)
2. Darbinyan, A., Pozniak, P., Darbinian, N., White, M.K., Khalili, K.: *Compartmentalized Neuronal Cultures*, pp. 147–152. Humana Press, Totowa (2013)
3. Debanne, D., Campanac, E., Bialowas, A., Carlier, E., Alcaraz, G.: Axon physiology. *Physiol. Rev.* **91**(2), 555–602 (2011)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
5. Fanti, Z., Elena Martinez-Perez, M., De-Miguel, F.F.: Neurongrowth, a software for automatic quantification of neurite and filopodial dynamics from time-lapse sequences of digital images. *Dev. Neurobiol.* **71**(10), 870–881 (2011)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
7. Ho, S.Y., Chao, C.Y., Huang, H.L., Chiu, T.W., Charoenkwan, P., Hwang, E.: NeurphologyJ: an automatic neuronal morphology quantification method and its application in pharmacological discovery. *BMC Bioinformatics* **12**(1), 230 (2011)
8. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105. Curran Associates, Inc. (2012). <http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf>
9. Lin, G., Milan, A., Shen, C., Reid, I.: RefineNet: multi-path refinement networks for high-resolution semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1925–1934 (2017)
10. Lingor, P., Koch, J.C., Tönges, L., Bähr, M.: Axonal degeneration as a therapeutic target in the CNS. *Cell Tissue Res.* **349**(1), 289–311 (2012)
11. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015
12. Meijering, E.: Neuron tracing in perspective. *Cytometry Part A* **77**(7), 693–704 (2010)
13. Meijering, E.: Cell segmentation: 50 years down the road [life sciences]. *IEEE Signal Process. Mag.* **29**(5), 140–145 (2012)
14. Mesbah, R., McCane, B., Mills, S.: Deep convolutional encoder-decoder for myelin and axon segmentation. In: 2016 International Conference on Image and Vision Computing New Zealand (IVCNZ), pp. 1–6. IEEE (2016)
15. Naito, T., Nagashima, Y., Taira, K., Uchio, N., Tsuji, S., Shimizu, J.: Identification and segmentation of myelinated nerve fibers in a cross-sectional optical microscopic image using a deep learning model. *J. Neurosci. Methods* **291**, 141–149 (2017)
16. Rapoport, D.H., Becker, T., Madany Mamlouk, A., Schicktzanz, S., Kruse, C.: A novel validation algorithm allows for automated cell tracking and the extraction of biologically meaningful parameters. *PLoS ONE* **6**(11), e27315 (2011)

17. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). [https://doi.org/10.1007/978-3-319-24574-4\\_28](https://doi.org/10.1007/978-3-319-24574-4_28)
18. Zaimi, A., Wabartha, M., Herman, V., Antonsanti, P.L., Perone, C.S., Cohen-Adad, J.: AxonDeepSeg: automatic axon and myelin segmentation from microscopy data using convolutional neural networks. *Sci. Rep.* **8**(1), 1–11 (2018)
19. Zhang, J., Dashtbozorg, B., Bekkers, E., Pluim, J.P.W., Duits, R., ter Haar Romeny, B.M.: Robust retinal vessel segmentation via locally adaptive derivative frames in orientation scores. *IEEE Trans. Med. Imaging* **35**(12), 2631–2644 (2016)
20. Zhou, Z., Kuo, H.C., Peng, H., Long, F.: DeepNeuron: an open deep learning toolbox for neuron tracing. *Brain Inform.* **5**(2), 1–9 (2018)
21. Zille, M., et al.: Ferroptosis in neurons and cancer cells is similar but differentially regulated by histone deacetylase inhibitors. *eNeuro* **6**(1), ENEURO.0263-18.2019 (2019). <https://doi.org/10.1523/ENEURO.0263-18.2019>