NEUROPHOTONICS

PERSPECTIVE

Addressing annotation and data scarcity when designing machine learning strategies for neurophotonics

Catherine Bouchard,[a,b,†] **Renaud Bernatchez**,[a,b,†] **and Flavie Lavoie-Cardinal**[a,b,c,*]
[a]CERVO Brain Research Centre, Québec, Québec, Canada
[b]Université Laval, Institute Intelligence and Data, Québec, Québec, Canada
[c]Université Laval, Département de psychiatrie et de neurosciences, Québec, Québec, Canada

ABSTRACT. Machine learning has revolutionized the way data are processed, allowing information to be extracted in a fraction of the time it would take an expert. In the field of neurophotonics, machine learning approaches are used to automatically detect and classify features of interest in complex images. One of the key challenges in applying machine learning methods to the field of neurophotonics is the scarcity of available data and the complexity associated with labeling them, which can limit the performance of data-driven algorithms. We present an overview of various strategies, such as weakly supervised learning, active learning, and domain adaptation that can be used to address the problem of labeled data scarcity in neurophotonics. We provide a comprehensive overview of the strengths and limitations of each approach and discuss their potential applications to bioimaging datasets. In addition, we highlight how different strategies can be combined to increase model performance on those datasets. The approaches we describe can help to improve the accessibility of machine learning-based analysis with limited number of annotated images for training and can enable researchers to extract more meaningful insights from small datasets.

Keywords: machine learning; active learning; weakly supervised learning; domain adaptation; image analysis

Paper 23033SSPER received May 3, 2023; revised Jul. 19, 2023; accepted Jul. 20, 2023; published Aug. 24, 2023.

1 Introduction

The recent emergence of machine learning approaches has transformed the landscape of biomedical data analysis. Since the first demonstration that the U-Net could be successfully applied to single cell segmentation on a limited number of training samples,[1] important efforts have been made in developing machine learning tools that are accessible to the bioimaging community.[2–5] Such tools include user interfaces that integrate image visualization, labeling, training, and prediction, such as Ilastik[5] and Napari,[6] as well as pre-trained algorithms designed to be easily applied to new data, such as Cellpose[4] and deepImageJ.[7] Efforts are also being made to facilitate access to the computer resources needed to train the models for microscopy image analysis, which are often a barrier, such as ZeroCostDL4Mic[2] and the BioImage Model Zoo.[3] However, supervised machine learning models require datasets that are specifically processed and annotated for the task that they are designed for (e.g., segmentation, detection, and

[*] Address all correspondence to Flavie Lavoie-Cardinal, flavie.lavoie-cardinal@cervo.ulaval.ca

[†] These authors contributed equally.

classification). While models can be pretrained on open-access bioimaging datasets,[3,8–10] a fine-tuning step with a subset of annotated data needs to be performed to adapt the model to a new bioimaging analysis task. The annotated training datasets need to be large enough to represent accurately the full data distribution to ensure that the model generalizes well at inference. When the model is applied to a new dataset or to a new batch of images, changes in the features defining the objects of interest may reduce the performance or require the model to be optimized. For biological experiments, this means that any change in the experimental settings may require re-annotation of a subset of the data and retraining or fine-tuning of the model to fit the new data distribution. In research fields, such as neurophotonics, in which data acquisition is costly and annotation requires trained experts, strategies need to be developed to mitigate annotation complexity and increase the robustness of machine learning models to data variability and data imbalance. To support the democratization of machine learning approaches for biomedical image analysis, sharing of open-source models and open-access datasets needs to be combined with optimized and simple annotation strategies accessible to domain experts. Here, we address two aspects that can potentially increase the accessibility to annotated data in neurophotonics: (1) labeling complexity and (2) data scarcity. We present methods we have applied to facilitate the application of machine learning to solve real neurophotonics-specific challenges we have encountered.

2 Discussion

2.1 Labeling Complexity

Training fully supervised machine learning models to perform analysis tasks on bioimaging datasets is challenging as it requires large amount of precisely annotated images.[2,11,12] Considering that the acquisition of new data or the annotations of the full dataset is not always possible due to ethical, time, or cost constraints, alternative strategies need to be proposed to democratize machine learning approaches in the field of neurophotonics.

We have recently addressed the challenge of labeling scarcity using weakly supervised learning approaches. We demonstrated that bounding boxes and binary annotations can replace precise contour annotations to train deep learning models on different tasks (instance segmentation, semantic segmentation, localization, and detection)[8,13] [Fig. 1(a)]. Simplifying the annotation process reduces both the annotation time and the inter-expert variability. It is a promising avenue in cases where the annotation task proves to be tedious and requires the involvement of trained experts.

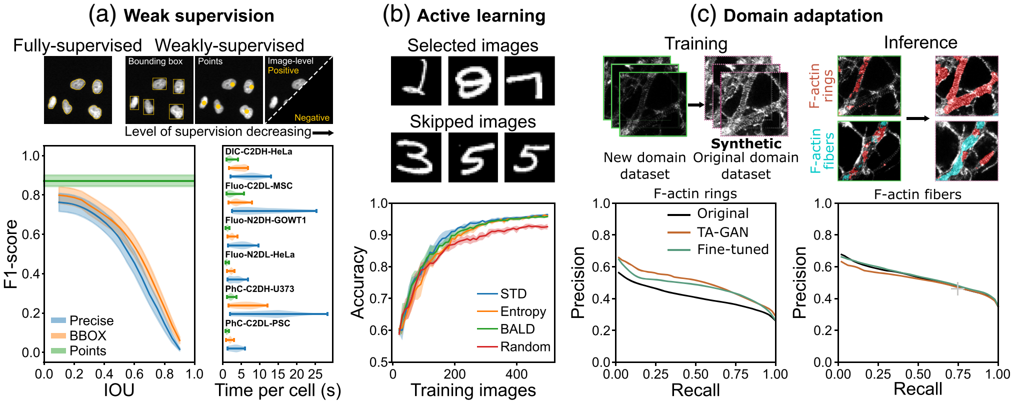

Fig. 1 Applications of weak supervision, active learning, and domain adaptation. (a) Similar performance is obtained for a segmentation task on the cell tracking challenge using weakly and fully supervised training schemes. Weak supervision significantly reduces the annotation time.[13] (b) Active learning allows to obtain a better classification performance with fewer training images compared to random selection of images.[14] (c) Domain adaptation can help apply a previously trained model to new images from a different domain (e.g., batch, device) by adapting the new images to the original domain.[15]

In some cases, the available datasets do not provide any or only a very limited number of annotated images, making the training of a deep neural network inconceivable. One promising approach to increase the number of annotated images within a larger dataset in an efficient way is active learning[16] [Fig. 1(b)]. In an active learning context, the training dataset is iteratively created by asking an expert to label the most informative samples, i.e., those expected to bring the most improvement to the model's performance, using a measure of prediction uncertainty[17,18] or sample diversity.[19,20] However, when the annotation cost (i.e., time to produce an annotation) of each sample is not constant across the dataset, it should be considered in the design of the active learning model to avoid increasing the total annotation cost while reducing the number of annotations.[14,21] We have shown how a trade-off between annotation cost and model performance can be achieved in a simple task, which could be extended to neurophotonics data.[14]

In neurophotonics datasets, researchers often face the challenge of defining a precise ground truth for a specific task (e.g., identifying object borders for a segmentation task). In this context, large intra- and inter-expert variability can be observed.[13] Crowdsourcing is a strategy used to gain access to a large pool of annotations from the community. Crowdsourcing helps alleviate the challenge of annotating large amounts of data by spreading the task across many people, also enabling the collection of multiple annotations for each sample. This is initially used to help aggregate the annotations from non-expert users, but it also grants access to a measure of the confidence on the obtained annotations. Such a measure could be leveraged to consider the commonly encountered annotation variability of biological structures in neurophotonics datasets.

Self-supervised learning[22,23] (SSL) is a promising avenue for addressing annotated data scarcity when large datasets are available but the accessibility of ground truth annotations is limited. SSL is a two-steps paradigm where (1) the general representation of the domain is learned using a pretext task that does not require labeled data, and (2) the downstream task is learned using the fraction of the dataset that is labeled.[24] The use of SSL for neurophotonics can be limited by the identification of reliable pretext tasks that are well adapted for microscopy images. Common pretext tasks for images include context prediction,[22,25] jigsaw puzzles,[26] rotation prediction,[27] and colorization;[28] all of which are not directly applicable to microscopy images. The context prediction and the jigsaw puzzle tasks could have more than one possible answer (particularly for images where structures are far apart), rotations are not defined in the plane of the image, and pseudo-colors are arbitrarily defined from photon-counts. Instance discrimination,[29] geometric self-distillation,[30] classification of image parameters (e.g., scale[31]), and image prediction[32,33] are all pretext tasks that are applicable to microscopy images, do not require semantic labels, and yet still enable the model to learn generalizable representations of the data. Image prediction can be used as a pretext task to learn denoising in temporal imaging data, improving signal-to-noise ratio in calcium[34,35] and voltage[36] imaging without the need for ground-truth denoised images, which are difficult to obtain. These methods take advantage of the spatial relationship between consecutive frames to learn to generate images with reduced noise. Such denoising approaches can be applied in real-time during imaging, proving particularly useful in photon-limited contexts, such as two-photon microscopy.[37]

2.2 Data Scarcity

The expanding range of microscopes and imaging systems that are routinely applied to neuroscience research questions unlocks new insights into complex biological processes. The gained flexibility in the choice of devices and acquisition parameters to characterize a given biological structure can lead to an increased variability in the properties of the generated datasets for a very similar research question. Images from the same structure acquired by two different groups or even at different time points on the same device can belong to different data distributions. Notably, models trained on a given type of images (dimension, resolution, and modality) will not necessarily be directly applicable to a new dataset.[13] While it can be obvious when addressing completely different modalities, it becomes problematic when differences between the image datasets are barely perceptible even for a trained expert [Fig. 1(c)].

A model proven to be effective for the segmentation of F-actin nanostructures in STimulated Emission Depletion (STED) images of fixed hippocampal neurons[8] was unsuccessful in segmenting the same structures on new images acquired a few years later on the same device. To avoid annotating a new dataset to retrain a deep neural network from scratch, we explored two alternative approaches: transfer learning (fine-tuning the original segmentation network) and synthetic data generation using a conditional generative adversarial network (cGAN) for domain adaptation.[15] cGANs create synthetic images from one domain based on input images from another domain.[38,39] It allows adaptation of the image features from a new distribution to match those of the original distribution [Fig. 1(c)]. Both methods (transfer learning and training on domain-adapted synthetic data) improved the segmentation accuracy on the new dataset over the original segmentation network. Synthetic data may also be generated using biologically and optically accurate simulations without the use of neural networks.[40,41] This proves useful in particularly data-hungry learning methods, such as reinforcement learning, where acquiring the required amount of imaging data for training is unreasonable.[42]

The idea of using knowledge obtained from one dataset before learning a task on a second dataset was proven effective for many applications and is a well-established method for addressing data scarcity.[43,44] When training a deep neural network using transfer learning, we rely on a large labeled dataset to learn the initial weights of the model. This model can be fine-tuned to the new domain associated with a smaller dataset (Fig. 2). Using transfer learning reduces the training cost since the pretraining step is performed on very large open-source datasets. However, transfer learning might offer little benefit in neurophotonics since the features defining the structures in microscopy images differ significantly from the features in the common large datasets composed of natural images.[44] Encouraging the shift toward open-source data will allow building huge field-specific community datasets from which general representations can be learned, similarly to commonly used computer vision datasets.

2.3 Data Imbalance

A dataset where the number of training examples is not constant across classes is called imbalanced. Data imbalance is a prevalent challenge across bio-imaging applications, because elements of interest tend to correlate with fewer occurrences.[45] In neurophotonics, segmentation tasks often meet the data imbalance problem since the number of pixels to which a class is assigned can be far inferior to either the number of pixels from a different class (e.g., cell bodies versus neuritis[46] and active versus inactive neurons in two-photon calcium imaging[47]) or the number of unlabeled pixels (background or non-studied structures). If this data imbalance is not addressed, models can become overconfident for the more prevalent classes and avoid learning the distinctive features of less common classes.

Data augmentation in computer vision is the process of applying geometric transformations (e.g., rotations, scaling, flips, crops, and skewing), color transformations (e.g., brightness, contrast, and saturation), or appearance transformations (e.g., Gaussian filtering, Sobel filtering, and noise addition) to increase the number of different training examples from a given number of data samples.[48] Data augmentation can alleviate the consequences of data imbalance by augmenting the images of rare classes until their number matches the most common class. However, not all types of transformations designed for natural images can be applied to microscopy images. For example, scaling images during training could prevent a model from learning

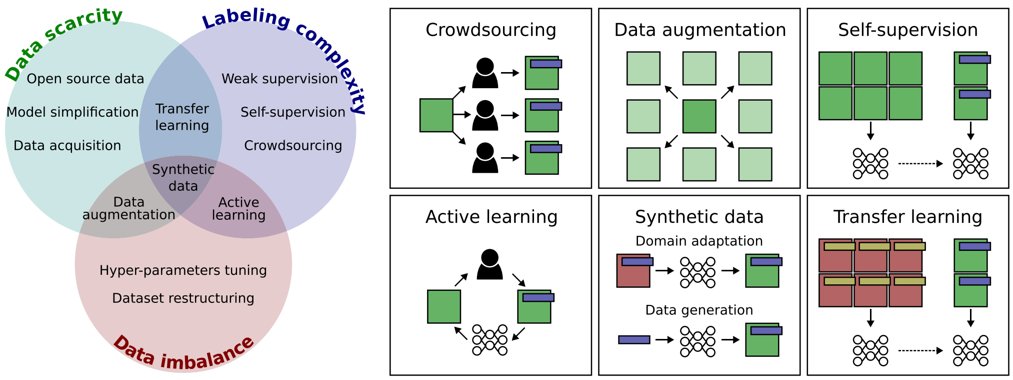

Fig. 2 (Left) Solutions to address common data-driven challenges in neurophotonics: data scarcity, label scarcity, and data imbalance. (Right) Green and red boxes represent images from different domains, and blue and yellow rectangles represent annotations for different tasks.

size features that can be useful for analyzing images acquired at a constant magnification.[49] Colors also hold a different meaning in fluorescence microscopy than for natural images and this meaning must be preserved through color transformations.[50] The application of filters can decrease the resolution of the images, and nanoscale elements can be lost through the blurring effects, affecting the possibility for the model to recognize nanoscale features.[51] Careful considerations must be taken for what types of transformations can be applied to the images without altering the significance of their assigned labels.

3 Conclusion

In the field of neurophotonics, challenges associated with data and label scarcity can be exacerbated by the complexity of the image acquisition, the requirement for expert knowledge for annotations, and the experimental variability. We covered a few possible methods for tackling data and label scarcity through concrete challenges we have encountered. To democratize machine learning-based quantitative bioimaging, the development of approaches that are accessible, reproducible, documented, and broadly available to the community will be essential. Their deployment will be coupled with strategies to improve the efficiency of the training process of deep learning models, both in terms of required data and annotations.

Disclosures

The authors have no conflicts of interest to declare.

Acknowledgments

Funding was provided by grants from the Natural Sciences and Engineering Research Council of Canada (NSERC) (Grant No. RGPIN-06704-2019); the Neuronex Initiative [National Science Foundation (Grant No. 2014862), Fond de recherche du Québec - Santé (Grant No. 295824)]; and the New Frontiers in Research Fund (Grant No. NFRFE-2020-00933). F.L.C. is a Canada Research Chair Tier II. C.B. is supported by scholarships from NSERC, from the Fonds de Recherche Nature et Technologie (FRQNT) Quebec, from the FRQNT strategic cluster UNIQUE, and by a Leadership and Scientific Engagement Award from Université Laval.

References

1. O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th Int. Conf.*, Munich, Germany, Proceedings, Part III 18, pp. 234–241, Springer International Publishing (2015).
2. L. von Chamier et al., "Democratising deep learning for microscopy with ZeroCostDL4Mic," *Nat. Commun.* **12**(1), 2276 (2021).
3. W. Ouyang et al., "Bioimage model zoo: a community-driven resource for accessible deep learning in bioimage analysis," bioRxiv, 2022–06 (2022).
4. C. Stringer et al., "Cellpose: a generalist algorithm for cellular segmentation," *Nat. Methods* **18**(1), 100–106 (2021).
5. S. Berg et al., "Ilastik: interactive machine learning for (bio)image analysis," *Nat. Methods* **16**(12), 1226–1232 (2019).
6. N. Sofroniew et al., *Napari: A Multi-Dimensional Image Viewer for Python*, Zenodo (2022).
7. E. Gómez-de Mariscal et al., "Deepimagej: a user-friendly environment to run deep learning models in imagej," *Nat. Methods* **18**(10), 1192–1195 (2021).
8. F. Lavoie-Cardinal et al., "Neuronal activity remodels the F-actin based submembrane lattice in dendrites but not axons of hippocampal neurons," *Sci. Rep.* **10**(1), 11960 (2020).
9. V. Ulman et al., et al., "An objective comparison of cell-tracking algorithms," *Nat. Methods* **14**(12), 1141–1152 (2017).
10. C. Spahn et al., "Deepbacs for multi-task bacterial image analysis using open-source deep learning approaches," *Commun. Biol.* **5**(1), 688 (2022).
11. H. Spiers et al., "Citizen science, cells and CNNs – deep learning for automatic segmentation of the nuclear envelope in electron microscopy data, trained with volunteer segmentations," (2020).
12. D. A. Van Valen et al., "Deep learning automates the quantitative analysis of individual cells in live-cell imaging experiments," *PLoS Comput. Biol.* **12**(11), e1005177 (2016).

13. A. Bilodeau et al, "Microscopy analysis neural network to solve detection, enumeration and segmentation from image-level annotations," *Nat. Mach. Intell.* **4**(5), 455–466 (2022).
14. R. Bernatchez, A. Durand, and F. Lavoie-Cardinal, "Annotation cost-sensitive deep active learning with limited data (student abstract)," *Proc. AAAI Conf. Artif. Intell.* **36**(11), 12913–12914 (2022).
15. C. Bouchard et al., "Resolution enhancement with a task-assisted GAN to guide optical nanoscopy image analysis and acquisition," *Nat. Mach. Intell.* (2023).
16. B. Settles, "Active learning," *Synth. Lect. Artif. Intell. Mach. Learn.* **6**(1), 1–114 (2012).
17. Y. Gal, R. Islam, and Z. Ghahramani, "Deep Bayesian active learning with image data," in *Int. Conf. Mach. Learn. (ICML)*, PMLR, pp. 1183–1192 (2017).
18. J. T. Ash et al., "Deep batch active learning by diverse, uncertain gradient lower bounds," in *Eighth Int. Conf. Learn. Represent.* (2020).
19. Y. Geifman and R. El-Yaniv, "Deep active learning over the long tail," arXiv:1711.00941 (2017).
20. C. Shui et al., "Deep active learning: unified and principled method for query and training," in *Int. Conf. Artif. Intell. and Statistics*, pp. 1308–1318, PMLR (2020, June).
21. B. Settles, M. Craven, and L. Friedland, "Active learning with real annotation costs," in *NIPS Workshop Cost-Sensitive Learn.*, p. 10 (2008).
22. C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *IEEE Int. Conf. Comput. Vision (ICCV)*, IEEE, Santiago, pp. 1422–1430 (2015).
23. L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: a survey," *IEEE Trans. Pattern Anal. Mach. Intell.* **43**(11), 4037–4058 (2020).
24. S. Albelwi, "Survey on self-supervised learning: auxiliary pretext tasks and contrastive learning methods in imaging," *Entropy (Basel)* **24**(4), 551 (2022).
25. D. Pathak et al., "Context encoders: feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 2536–2544 (2016).
26. M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," arXiv:1603.09246 (2017).
27. S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," in *Int. Conf. Learn. Represent. (ICLR)*, OpenReview, Vancouver, BC, Canada (2018).
28. G. Larsson, M. Maire, and G. Shakhnarovich, "Colorization as a proxy task for visual understanding," in *IEEE Conf. Comput. Vision and Pattern Recognit. (CVPR)*, IEEE, Honolulu, Hawaii, pp. 840–849 (2017).
29. Y. Tu et al., "SIFLoc: a self-supervised pre-training method for enhancing the recognition of protein subcellular localization in immunofluorescence microscopic images," *Brief Bioinf.* **23**(2), bbab605 (2022).
30. B. Midtvedt et al., "Single-shot self-supervised object detection in microscopy," *Nat Commun* **13**(1), 7492 (2022).
31. C. Krug and K. Rohr, "Unsupervised cell segmentation in fluorescence microscopy images via self-supervised learning," *Lect. Notes Comput. Sci.* **13363**, 236–247 (2022).
32. A. X. Lu et al., "Learning unsupervised feature representations for single cell microscopy images with paired cell inpainting," *PLoS Comput. Biol.* **15**(9), e1007348 (2019).
33. J. Ma et al., "3D nucleus instance segmentation for whole-brain microscopy images," *Lect. Notes Comput. Sci.* **12729**, 504–516 (2021).
34. J. Lecoq et al., "Removing independent noise in systems neuroscience data using deepinterpolation," *Nat. Methods* **18**, 1401–1408 (2021).
35. X. Li et al., "Reinforcing neuron extraction and spike inference in calcium imaging using deep self-supervised denoising," *Nat. Methods* **18**, 1395–1400 (2021).
36. J. Platisa et al., "High-speed low-light in vivo two-photon voltage imaging of large neuronal populations," *Nat. Methods* **20**, 1095–1103 (2023).
37. X. Li et al., "Real-time denoising enables high-sensitivity fluorescence time-lapse imaging beyond the shot-noise limit," *Nat. Biotechnol.* **41**, 282–292 (2023).
38. M. Mirza and S. Osindero, "Conditional generative adversarial nets," arXiv:1411.1784 (2014).
39. P. Isola et al., "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 1125–1134 (2017).
40. A. Song et al., "Neural anatomy and optical microscopy (NAOMi) simulation for evaluating calcium imaging methods," *J. Neurosci. Methods* **358**, 109173 (2021).
41. Y. Zhang et al., "Rapid detection of neurons in widefield calcium imaging datasets after training with synthetic data," *Nat. Methods* **20**, 747–754 (2023).
42. B. Turcotte et al., "pySTED: a STED microscopy simulation tool for machine learning training," in *AAAI Workshop on AI to Accel. Sci. and Eng. (AI2ASE)* (2022).
43. M. L. Hutchinson et al., "Overcoming data scarcity with transfer learning," arXiv:1711.05099 (2017).
44. M. Raghu et al., "Transfusion: understanding transfer learning for medical imaging," in *Adv. Neural Inf. Process. Syst. 32*, Curran Associates, Inc. (2019).

45. L. Gao et al., “Handling imbalanced medical image data: a deep-learning-based one-class classification approach,” *Artif. Intell. Med.* **108**, 101935 (2020).
46. E. M. Gil et al., “Comparing the segmentation of quantitative phase images of neurons using convolutional neural networks trained on simulated and augmented imagery,” *Neurophotonics* **10**(3), 035004 (2023).
47. S. Soltanian-Zadeh et al., “Fast and robust active neuron segmentation in two-photon calcium imaging using spatiotemporal deep learning,” *Proc. Natl. Acad. Sci. U. S. A.* **116**(17), 8554–8563 (2019).
48. C. Shorten and T. M. Khoshgoftaar, “A survey on image data augmentation for deep learning,” *J. Big Data* **6**(1), 1–48 (2019).
49. J. Ma et al., “Review of image augmentation used in deep learning-based material microscopic image segmentation,” *Appl. Sci.* **13**(11), 6478 (2023).
50. W. D. Cameron et al., “Leveraging multimodal microscopy to optimize deep learning models for cell segmentation,” *APL Bioeng.* **5**(1), 016101 (2021).
51. I. Groves et al., “Bespoke data augmentation and network construction enable image classification on small microscopy datasets,” bioRxiv (2022).

Catherine Bouchard is a PhD student in electrical engineering at Université Laval. She is interested in the development of deep learning algorithms integrated into super-resolution microscopy acquisition loops.

Renaud Bernatchez completed a master’s degree in biophotonics in 2023 and was studying the development of supervised learning algorithms with little annotated data.

Flavie Lavoie-Cardinal is an associate professor in the Department of Psychiatry and Neuroscience at Université Laval, principal investigator at the CERVO Brain Research Center and director of the Health and Life Science Research Axis at the Institute for Intelligence and Data. Her team is working on the development of smart microscopy approaches applied to neuroscience.