Denoising Tissue-scatter Limited Deep Cortical Image Using U-Net

Benny Weng, and Alex Kiral

Abstract—In order to establish a complete understanding of the neocortex, neuroscientists must be able to perform imaging at depths beyond $750\ \mu m$, which is beyond the capability of microscopy setups today. The fundamental challenge is the emergence of a substantial background signal at the lowest layer of the neocortex due to scattering from shallower tissue into the plane-of-interest. In this paper, we propose a novel computational and experimental procedure for the removal of this background. We utilize a U-net architecture to learn the spatial structure of the background signal given images with both background and signal. Using realistic simulation data, we find that our method can improve Pearson correlation coefficients of the extract neuronal time traces with the ground truth from 0.85 to 0.98. In addition, we propose an experimental calibration methodology which will allow for the training of the model in the laboratory, by utilizing background-free images at shallower layers.

Index Terms—Neural imaging, Denoiser, U-Net

1 Introduction

TWO-PHOTON microscopy [1] is considered a workhorse in modern neuroscience due to its superior optical sectioning capability (subcellular resolution both laterally and axially) and increased scattering length resulting from the utilization of near infrared (NIR) excitation sources. The application of two-photon excitation in scanning microscopes has enabled routine *in vivo* imaging in the neocortices of rodents performing behavioral tasks.

The mammalian neocortex is often considered to have six laminar layers, demarcated by anatomical features but also populated by neurons belonging to different genetic, projection, and functional types. The cortical column is often considered a basic architecture of cortical computation, but how the six layers work in concert remains poorly understood. While two-photon excitation can easily access layer 2/3 and in a well-engineered system, down to layer 5, imaging layer 6 ($> 750\ \mu$m) in the rodent neocortex remains challenging. Given the anisotropic nature of the brain, most of the ballistic photons are scattered before reaching the intended depth, creating a background distinct in both spatial and temporal scales and drowning out the neural signals. This is largely why people have refrained from imaging these depths, even though there is evidence obtained through alternative non-imaging modalities suggesting that layer 6 neurons are responsible for modulating signal amplitudes [2], which is very distinct from the roles played by other layers.

Strategies for accessing layer 6 in the rodent neocortex have largely involved novel experimental techniques, such as single-pulse-per-pixel excitation and three-photon excitation. These techniques often require expensive bespoke femtosecond laser sources, and are challenging to set up both electronically and optically. What remains relatively under-explored are computational strategies.

- *B. Weng is with the Department of Applied Physics, Stanford University, Stanford, CA, 94305. E-mail: bweng914@stanford.edu*
- *A. Kiral is with the Department of Physics, Stanford University, Stanford, CA, 94305. E-mail: akiral@stanford.edu*

2 Related Work

A number of neural network-based denoising approaches exist in the literature, but they are generally based on Noise2Noise [3] or its derivative Noise2Self [4], [5], [6], which are intended for mean-zero shot noise removal. These are used with success on data from shallower layers, but they fall short on data with significant background. This is most likely because the large fluorescence background as a result of tissue scattering is not mean-zero and is generally slower temporally and larger-scale spatially. The scattered background also comes with its own mean-zero shot noise component.

3 Proposed Method

In this paper, we introduce a procedure for supervised learning of a U-Net CNN for background removal in deep cortical imaging. The U-net architecture, first introduced in 2015 for biological image segmentation, is an efficient way to construct a convolutional neural network for image processing. In this architecture, shown schematically in Fig. 1, We use a symmetric structure, which repeatedly down-samples the image to a lower spatial resolution, followed by up-sampling back to the original spatial resolution. This process allows the network to learn larger scale spatial structures which are important in images. A U-net also contains skip connection, which directly connects every resolution in the initial down-sampling phase with the layer in the up-sampling phase with the same resolution. This technique enables the network to retain information on smaller spatial structures, which would have been lost in the down-sampling procedure.

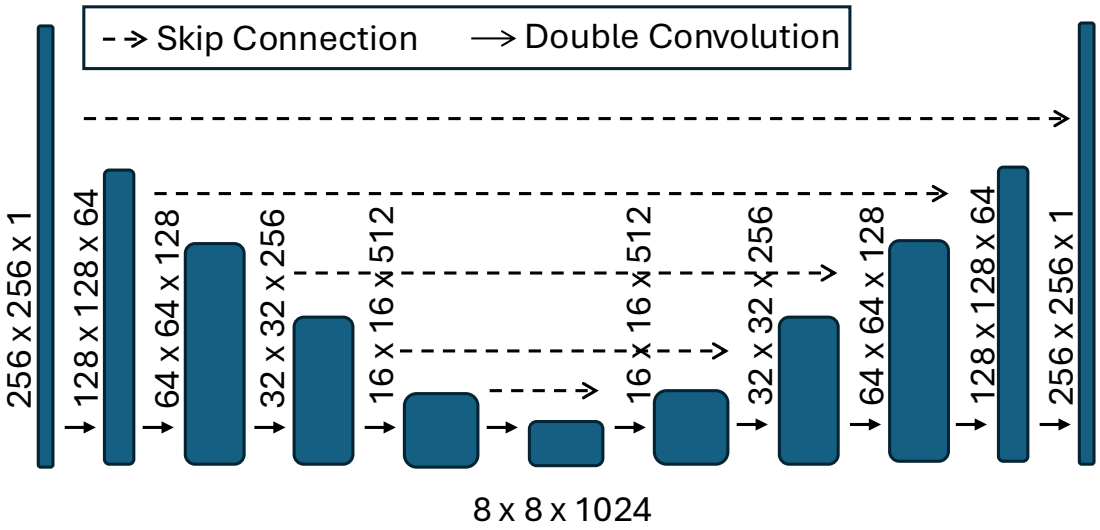

Fig. 1. A schematic diagram of the U-Net architecture used to model the background in neuronal activity images. A measured or simulated background + signal intensity image of size 256×256 is fed in as input on the left. At each step in the down-sampling, we perform double convolution, which consists of 3×3 convolutional layers, a batch normalization layer, and then a ReLU layer, performed twice. There are n convolutional layers applied to the input, where n is also the number of channels of the next block. After this a pooling operation is performed to reduce the spatial size of the image. In the up-sampling process, we perform a convolutional transpose in order to increase the spatial size and decrease the number of channels of the image. After that we again perform double convolution. Finally, skip connections are connected between all blocks of the same size.

3.1 Loss Function

We use an MSE loss to train the model to identify the background given an image with both signal and background as input. This has the form

$$\mathcal{L}(\mathbf{y};\theta) = \frac{1}{N}\sum_{i=1}^{N} ||(y_i - z_i) - h_\theta(y_i)||_2^2 \tag{1}$$

where y_i is the true total (i.e. background + signal), z_i is the true signal, and $h_\theta(y_i)$ is the model prediction with parameters θ. N represents the number of images in the batch, and we use a norm to indicate that this is a pixel-wise computation over a whole image. This procedure requires both the true total image and true signal, which is trivial to generate when working with simulation data, but in experimental setups requires a calibration procedure described in Section 3.2. Other loss functions are possible with the same methodology, and the optimal choice may depend on the post-processing that is done to extract neuronal signals from the background-subtracted movies. In particular, L1 loss may be an attractive choice due to the spatial sparsity of the signal.

Note that we choose to learn the background rather than directly estimating signal from the true signal. We made this choice because we believe that the background has larger spatial correlations than the signal, which a convolutional neural network may be more proficient at learning. Based on preliminary investigation, this spatial structure seems to exist in the experimental background as well, but if this is not true then it may be more fruitful to train the neural network to learn the signal, rather than background.

3.2 Experimental Calibration

Acquiring separate background + signal and in-focus movies simultaneously in simulation is trivial, but doing so in a real experimental setting is less straightforward. This will be crucial for the training and background estimation of real experimental data, because while the general structure of the experimental data can be simulated computationally, different microscopes, neuronal labeling strategies, light sources, and other experimental conditions can result in differences in signal and background distributions and dynamics.

In this section, we propose a simple experimental schematic intended for collecting a clean in-focus movie and the same movie with significant background simultaneously. We plan to construct an optical setup where two beams are focused on the same imaging plane in a shallower layer, but one of the beam paths will have its objective back aperture plane conjugated to an adaptive optical element, such as a spatial-light modulator (SLM), with which the generation of arbitrary 3D hologram is possible. We will use the SLM to (1) generate a suboptimal focal point spread function (PSF) for the target imaging plane to mimic the corruption of the PSF due to the anisotropy of brain tissue, and (2) generate either a series of weaker focal points axially or an elongated PSF such as a Bessel beam to increase the excitation of out-of-focus tissue and increase the background at the target imaging plane. We will then construct an optical delay line such that the time of arrival of the two beams will be separated by tens of nanosecond, long enough such that we can gate the detector to collect the signals from the two excitation paths separately, but short enough that it is significantly faster than the typical time scale of calcium dynamics (~ 10 ms) so that the two movies can be considered simultaneous from a biological point of view. These two movies will then be used for training the background-subtraction model.

This experimental calibration procedure should in principle be highly flexible so that the user can explore a variety of PSF corruption and background conditions and generate

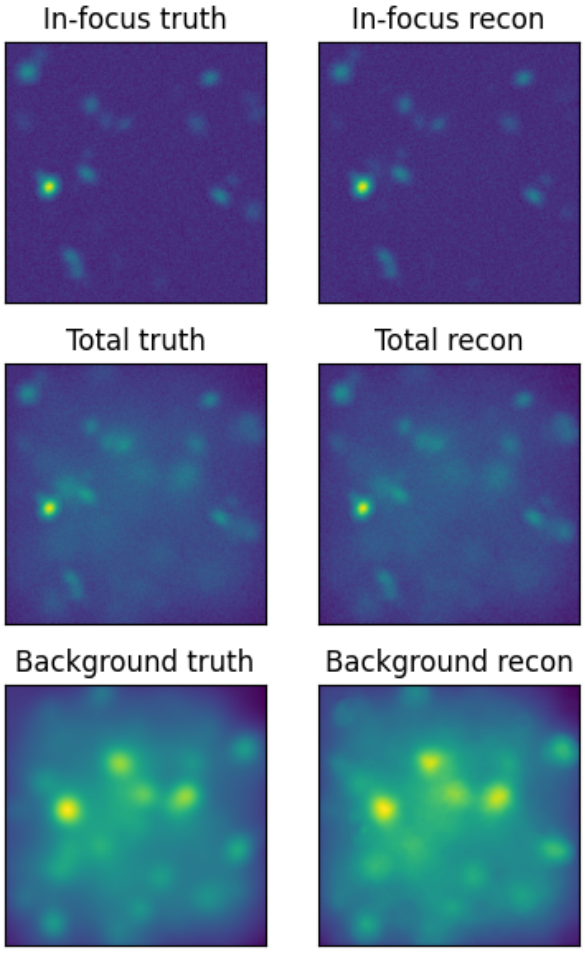

Fig. 2. A single frame in the movie for the in-focus plane (i.e. the plane of interest), the total, and the background for both the ground truth and the reconstruction. Ground truth is generated through simulation, described in Section 4.1. The total truth and total reconstructed are the same image, the background reconstruction is the neural network output after feeding in the total image, and the in-focus reconstruction is the total minus the background reconstruction.

a variety of models based on these contexts. However, we should point out that this procedure relies on the fact that the data acquisition speed is faster than the excitation laser repetition rate so that the dead time between pulses can be leveraged [7].

4 Experimental Results

In this paper, we train and test our model on simulated movies of neuronal activity with scattering of light between different imaging depths. Our simulation procedure consists of creating signal + background movies congruent with photon physics and optical principles. We show relevant metrics for the training of the network, and biologically relevant quantitative results on the test datasets.

4.1 Simulating neuronal videos

The image formation in deep tissue can be represented as follows:

$$A(D) = GT(D) * \mathrm{PSF}(D) + \sum_{d=0}^{D} e^{-\frac{D-d}{l_s}} \times GT(d) * g(\sigma(d)) + \sqrt{\sum_{d=0}^{D} \lambda^2(d) + \lambda_{\mathrm{GT}}^2}, \tag{2}$$

where $A(D)$ is the formed image at imaging depth D, GT is the ground truth image, $PSF(D)$ is the microscope point spread function (PSF) at imaging depth D, λ_{GT} is the shot noise corresponding to the target imaging plane at depth D, $\lambda(d)$ is the shot noise corresponding to the plane at depth d, $GT(d)$ is the ground truth image at depth d, and $g(\sigma(d))$ is the Gaussian convolution kernel at depth d for the out-of-focus excitation beam with width $\sigma(d)$, which is determined by the objective lens used for imaging. We have included an exponential decay term modulated by the scattering length l_s to scale the intensity contribution of out-of-focus imaging planes to the background of the target imaging plane.

Here are some figure of merits for mouse neurons and calcium neuronal activities. Neurons are typically $\sim 10\ \mu m$ in diameter, and calcium spiking events start with a fast step-function-like onset followed by an exponential decay with $\tau \sim 1\,\mathrm{s}$; sequences of these events can be seen in Fig. 4. A typical two-photon scanning microscope has 30 Hz frame acquisition rate and roughly $1\ \mu m$ lateral resolution and $8\ \mu m$ axial resolution. We generated videos with $256 \times 256 \times 5000$ pixels, which correspond to fields-of-view of $256 \times 256\ \mu m^2$ and $\sim 3\ \mathrm{min}$ in length. We have randomly chosen cell diameters between 8 and 14 pixels, and set the calcium time constant to be $1.5\ \mathrm{s}$. We have also added Gaussian noise to our videos congruent with typical values in photo-multiplier tube based data collection setups. Given that mouse cortical layer 6 typically starts at $\sim\ 800\ \mu m$, we have sampled the depths d leading up to this target depth D with $8\ \mu m$ intervals, consistent with the smallest cell diameters we have chosen for our simulation as well as the typical axial-PSF size. This means we would sample 100 out-of-focus imaging planes for background generation.

For each simulated dataset, we created three videos: the in-focus imaging plane of interest, the background, and the sum of these two terms.

4.2 Training and testing

To train the model, we used a batch size of 40 and a 30-epoch training scheme with different learning rates (10^{-2} for the first ten epochs, 10^{-3} for the second ten epochs, and 10^{-4} for the last ten epochs). In Fig. 3, we plot the training loss against the batch number and show that the amount of training is sufficient.

We run the model on different simulated datasets and show pairs of ground truth and reconstructed example images in Fig. 2. Reconstructed backgrounds show excellent agreement with the ground-truth backgrounds with little sign of crosstalk with the in-focus imaging plane.

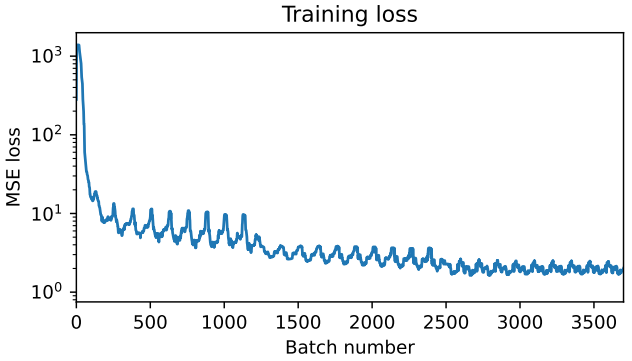

Fig. 3. Training loss as a function of batch number. Training takes place over 30 epochs, for a 5000 frame movie, with batch sizes of 40 frames. This gives a total of 3750 batches. We train at a learning rate of 10^{-2}, 10^{-3}, and 10^{-4} for 10 epochs each, which can be seen in the structure of the learning rate.

The most important consideration in the reconstruction of these neuronal movies is that the result does not quantitatively change the neuro-scientific conclusions. To quantitatively evaluate the goodness of the reconstruction, we use a robust-estimation-based approach [8] to extract neuronal time traces and compare them to the ground truth time traces we used to generate the simulation datasets. In Fig. 4, we plot the time traces extracted from the raw movie (left, red) and from the reconstructed in-focus movie (right, blue) and overlay them on the ground truth time traces (black). We show that even state-of-the-art cell extraction algorithms are not robust to fluctuations in a dominant background, whereas the time traces extracted from the reconstructed movie with learned-background subtraction show excellent agreement with ground truth. In Fig. 5, we computed the Pearson correlation coefficient between the extracted traces and the ground truth traces. The time traces extracted from the background-removed movie have $\rho = 0.98 \pm 0.001$ versus $\rho = 0.85 \pm 0.07$ from the original movie.

In addition, we include statistics for the PSNR as calculated on the 5000 frames of the test dataset. We see that both the background and signal (called plane-of-interest in the table) PSNR are both extremely good, with little variation in quality between different frames.

PSNR for Test Dataset

	Mean	Standard Deviation
Plane-of-interest	33.3	3.9
Background	32.6	0.3

5 Future Works

We believe that, in addition to the background which we have addressed in this work, scattering from earlier depths causes PSF distortion at later depths, due to the interference from scattered light on the main imaging beam. We can include this in our simulation by modifying the coefficients of each image in a Zernike polynomial expansion. This effect should be applied to every imaging plane, both in-focus and out-of-focus, although the effect on shallower layers will be less than on deeper layers. This effect should be applied randomly, and may change with the time-scale of the dynamics in the background.

In addition, we believe that it is feasible to implement an architecture similar to Noise2Self [4] in the existing developed framework. Noise2Self is a self-supervised machine learning for remove of mean-zero shot noise, which is a feature of both the background and signal discussed in this paper. It does this by utilizing the fact that the real images in neuron dynamics vary on a temporal scale which is much larger than the frame rate, so that successive frames, after removing noise, are highly correlated. This has been demonstrated to be highly successful for movies with low background, but initial experimentation with high background movies shows an effect in which an out-of-the-box Noise2Self model [5] causes the region-of-interest in the signal to lose definition and become indistinguishable from background. In order to fix this, we must train our noise-removal model to distinguish between the noise and background, which can be done by training the model in blocks of neighboring frames, and then performing the background removal algorithm demonstrated in this paper.

Experimental validation of our methodology is crucial for demonstrating its usefulness in real-world setups. Our simulations rely on assumptions about the structure of the background which may not be true in the laboratory. A crucial process in performing validation will be to perform background subtraction at a range of depths, starting with layer 5 data which has a small background, but where the signal is still fundamentally identifiable, and moving to layer 6 where signal to background ratio becomes far smaller.

6 Conclusion

In this paper, we have created a novel method for background removal intended for two-photon imaging of the deep cortical layers in mammalian brains. Based on a U-Net architecture, our method can effectively identify the scattered background, leverage its distinct spatiotemporal properties, and significantly improve the cell extraction fidelity. We have also proposed an experimentally viable approach to use this method for background subtraction of real neuronal data.

Deep layer 5 and layer 6 in mice neocortex are understudied due to the lack of optical access, and currently available experimental techniques require costly and complex experimental setups. Our method will allow the typical neuroscience lab with a standard two-photon microscope to gain access to deep cortical areas and discover new cortical neuronal population dynamics.

Acknowledgments

The authors would like to thank Dr. Gordon Wetzstein and Sonia Kim for their useful guidance in EE 367. BW thanks Dr. Mark Schnitzer for useful discussion and experimental guidance.

References

[1] W. Denk, J. H. Strickler, and W. W. Webb, "Two-photon laser scanning fluorescence microscopy," *Science*, vol. 248, no. 4951, pp. 73–76, 1990. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.2321027

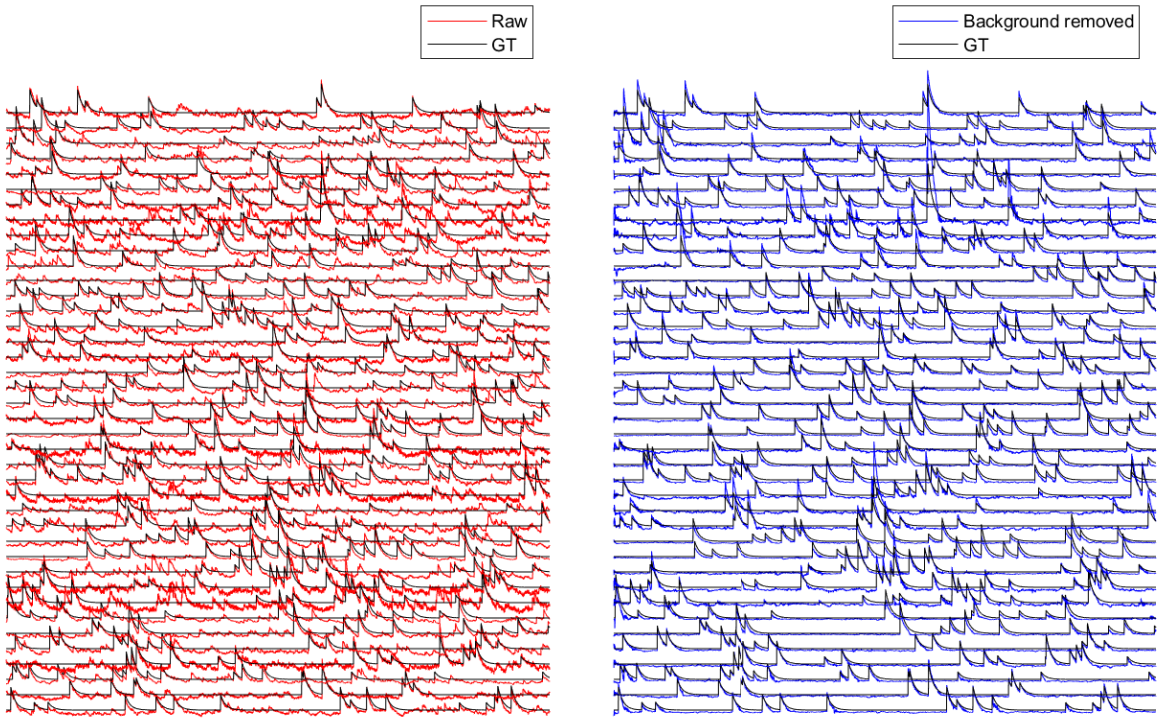

Fig. 4. Ground truth time traces *vs.* left: traces extracted from the raw movies and right: traces extracted from the background-subtracted movies. Traces of the same cells are aligned between the left and the right panels. The x-axis corresponds to 5,000 frames, and the y-axis corresponds to the change in intensity ($\Delta F/F$). A robust-estimation-based algorithm (EXTRACT [8]) was used for time trace extraction, seeded by the spatial filters used for creating the simulated movies.

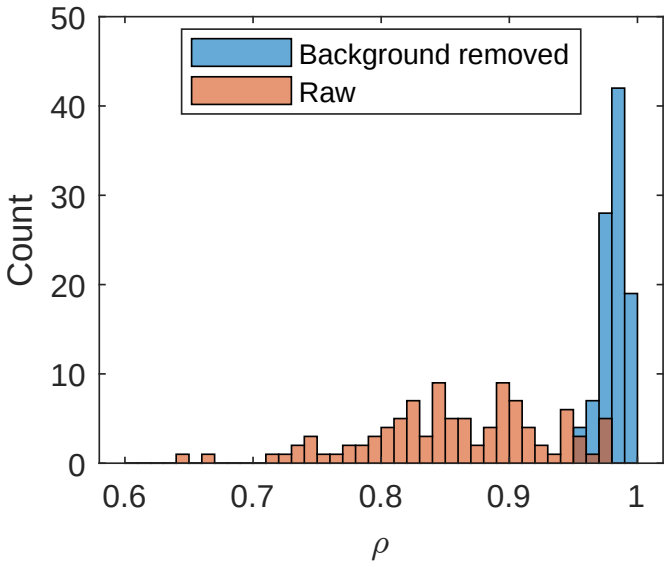

Fig. 5. Distribution of Pearson correlation coefficient between time traces extracted from the background-subtracted movie and ground truth time traces (blue), and between those from the raw movie and the ground truth. For the processed movie, $\rho = 0.98 \pm 0.001$ and for the original movie, $\rho = 0.85 \pm 0.07$.

[2] M. Vélez-Fort, E. F. Bracey, S. Keshavarzi, C. V. Rousseau, L. Cossell, S. C. Lenzi, M. Strom, and T. W. Margrie, "A circuit for integration of head- and visual-motion signals in layer 6 of mouse primary visual cortex," *Neuron*, vol. 98, no. 1, pp. 179–191.e6, 2026/02/19 2018. [Online]. Available: https://doi.org/10.1016/j.neuron.2018.02.023

[3] J. Lehtinen, J. Munkberg, J. Hasselgren, S. Laine, T. Karras, M. Aittala, and T. Aila, "Noise2noise: Learning image restoration without clean data," 2018. [Online]. Available: https://arxiv.org/abs/1803.04189

[4] J. Batson and L. Royer, "Noise2self: Blind denoising by self-supervision," 2019. [Online]. Available: https://arxiv.org/abs/1901.11365

[5] X. Li, G. Zhang, J. Wu, Y. Zhang, Z. Zhao, X. Lin, H. Qiao, H. Xie, H. Wang, L. Fang, and Q. Dai, "Reinforcing neuron extraction and spike inference in calcium imaging using deep self-supervised denoising," *Nature Methods*, vol. 18, no. 11, pp. 1395–1400, 2021. [Online]. Available: https://doi.org/10.1038/s41592-021-01225-0

[6] J. Lecoq, M. Oliver, J. H. Siegle, N. Orlova, P. Ledochowitsch, and C. Koch, "Removing independent noise in systems neuroscience data using deepinterpolation," *Nature Methods*, vol. 18, no. 11, pp. 1401–1408, 2021. [Online]. Available: https://doi.org/10.1038/s41592-021-01285-2

[7] J. Demas, J. Manley, F. Tejera, K. Barber, H. Kim, F. M. Traub, B. Chen, and A. Vaziri, "High-speed, cortex-wide volumetric recording of neuroactivity at cellular resolution using light beads microscopy," *Nature Methods*, vol. 18, no. 9, pp. 1103–1111, 2021. [Online]. Available: https://doi.org/10.1038/s41592-021-01239-8

[8] F. Dinç, H. Inan, O. Hernandez, C. Schmuckermair, O. Hazon, T. Tasci, B. O. Ahanonu, Y. Zhang, J. Lecoq, S. Haziza, M. J. Wagner, M. A. Erdogdu, and M. J. Schnitzer, "Fast, scalable, and statistically robust cell extraction from large-scale neural calcium imaging datasets," *bioRxiv*, 2024. [Online]. Available: https://www.biorxiv.org/content/early/2024/08/17/2021.03.24.436279