HHS Public Access

Author manuscript

Nat Mach Intell. Author manuscript.

An end-to-end recurrent compressed sensing method to denoise, detect and demix calcium imaging data

Kangning Zhang[1], **Sean Tang**[1,2], **Vivian Zhu**[1,3], **Majd Barchini**[1], **Weijian Yang**[1,*]

[1]Department of Electrical and Computer Engineering, University of California, Davis, CA 95616, USA

[2]Foothill High School, Pleasanton, CA 94588, USA

[3]Vista Del Lago High School, Folsom, CA 95630, USA

Abstract

Two-photon calcium imaging provides large-scale recordings of neuronal activities at cellular resolution. A robust, automated and high-speed pipeline to simultaneously segment the spatial footprints of neurons and extract their temporal activity traces while decontaminating them from background, noise and overlapping neurons is highly desirable to analyze calcium imaging data. In this paper, we demonstrate DeepCaImX, an end-to-end deep learning method based on an iterative shrinkage-thresholding algorithm and a long-short-term-memory neural network to achieve the above goals altogether at a very high speed and without any manually tuned hyper-parameter. DeepCaImX is a multi-task, multi-class and multi-label segmentation method composed of a compressed-sensing-inspired neural network with a recurrent layer and fully connected layers. It represents the first neural network that can simultaneously generate accurate neuronal footprints and extract clean neuronal activity traces from calcium imaging data. We trained the neural network with simulated datasets and benchmarked it against existing state-of-the-art methods with in vivo experimental data. DeepCaImX outperforms existing methods in the quality of segmentation and temporal trace extraction as well as processing speed. DeepCaImX is highly scalable and will benefit the analysis of mesoscale calcium imaging.

Two-photon calcium imaging can record neuronal activity at high resolution in deep brain tissue, and has been a workhorse in neuroscience to investigate neural circuits over the last two decades[1–5]. Recent advance in high-throughput two-photon microscopy enables the simultaneous high-speed recording of hundreds of thousands of neurons[6–14], underscoring the needs for an automated method to efficiently identify neuronal regions of interest (ROIs) and extract their fluorescence activities with high fidelity. The four critical tasks

* wejyang@ucdavis.edu .

Author Contributions

W.Y. and K.Z. conceived and designed this project. K.Z. developed and implemented the code for DeepCaImX. K.Z., S.T., V.Z. and M.B. implemented the code for other algorithms for comparison. K.Z. and W.Y. processed and analyzed the data. K.Z. and W.Y. wrote the paper. W.Y. supervised this research and provided the resources.

Competing Interests Statement

The authors declare no competing interests.

Code availability

Code for DeepCaImX, written in Python 3.9, can be accessed at DOI: 10.5281/zenodo.12650420[51].

in processing calcium imaging data are motion correction, denoising, segmentation, and temporal signal extraction. While motion correction has seen significant advancements with several effective algorithms[15–18], the methods for the other tasks[19–29] often fall short by being slow, requiring manual tuning, or having a limited fidelity in the output. Currently, there has not been a fully automated method that can perform denoising, segmentation and signal extraction with high speed and high quality.

Existing methods to process calcium imaging data can be categorized into model-based iterative optimization[19–21,23,24] and neural-network-based learning[26–29] algorithms. The former, such as principal component analysis and independent component analysis (PCA/ICA)[19], Suite2p[20], and CNMF[21] (constrained non-negative matrix factorization) or CaImAn[24], uses matrix factorization to model the imaging data with features of neuronal shapes and temporal dynamics. These methods generate the spatial footprint of the ROIs and demix their temporal traces simultaneously. They typically require lengthy processing and manual hyper-parameter tuning for satisfactory outcomes. Neural-network-based algorithms, such as STNeuroNet[26], SUNS[27], CITE-On[28], and DeepWonder[29], were developed to segment the calcium imaging data by learning from training dataset. Once trained, they are generally fast to operate. However, they require additional, model-based learning algorithms to extract and demix the neural activity traces. Self-supervised deep learning such as DeepCAD[30,31] and DeepInterpolation[32] were developed to denoise two-photon calcium imaging data, but they require other methods for the spatiotemporal analysis. So far, there has not been a single approach to simultaneously denoise and output both the spatial footprints and temporal activity traces.

We developed DeepCaImX, a fast, accurate, user-friendly and fully-automated end-to-end method to segment the neuronal cell bodies and extract, decontaminate and denoise their temporal activity traces for two-photon calcium imaging data. This supervised learning model employs a physics-aware, explainable and multi-task neural network. It is highly scalable to data size, and does not require any pre-processing such as spatial or temporal filtering, nor post-processing such as fine tuning or merging the ROIs, nor any hyper-parameter tuning. Our method begins with ISTA-Net[33], a compressed-sensing-inspired network, to denoise the video, remove the neuropil background and generate a spatially sparse representation of the neurons. The denoised and background-suppressed video is then recurrently processed by a long-short-term memory[34] (LSTM) layer in temporal domain, which learns the autoregressive model, and generates a time-series attention map for traces extraction in the subsequent 1D fully connected layers. These individual components, though distinct, work in concert and trained as a whole in an iterative (instead of sequential) manner. Such a model promotes a holistic optimization and outputs the segmentations and activity traces together, where signals from overlapping ROIs can be well demixed.

We evaluated DeepCaImX against top calcium imaging processing algorithms such as CaImAn[24], Suite2p[20], FISSA[23], SUNS[27], STNeuroNet[26], CITE-On[28], DeepWonder[29] and DeepCAD[30]-assisted CaImAn, in both simulation and in vivo experimental datasets. DeepCaImX outperforms these counterparts in the quality of segmentation and neural activity traces, while operating at a high speed. Its high speed, robustness and scalability

make it a powerful tool for analyzing large-scale calcium imaging data from mesoscale to 3D microscopes, enabling advanced studies of large-scale neuronal circuits.

Results

Principle of the end-to-end DeepCalmX model

Our model aims to predict the spatial footprint of neurons and extract their clean activity traces simultaneously, and address the challenges of neuropil background removal and signal demixing among overlapping neurons. A multi-task framework[35] could be effective and efficient for these goals, as the tasks of background and noise removal, ROIs segmentation, and traces extraction are coupled together through the sparse spatiotemporal dynamics of neuronal activity. We designed a single model to learn a generalized sparse representation of the data and branched it into spatial and temporal analyses. We trained the entire network on realistic calcium recording data simulated by the well-established NAOMi[36] method (Methods).

Our network is comprised of three highly explainable modules (Fig. 1, Extended Data Fig. 1). In the first module, we use an ISTA-Net[33] to suppress the neuropil background and noise frame-by-frame. The neuropil background generally has a smooth and spatially extended profile, whereas neuronal cell bodies have well defined spatial boundaries. ISTA-Net, a compressed-sensing-inspired neural network based on Iterative Shrinkage-Thresholding Algorithm (ISTA)[37], operates through multiple phases. In each phase, it converts the video from the spatial domain to a sparse representation, apply a soft threshold, and then revert back to the spatial domain (Methods). In the sparse representation, the neuropil has a diminished strength, and is gradually suppressed by the shrinkage threshold in each phase whereas the neurons with well-defined boundaries are preserved (Supplementary Fig. 1-2; Supplementary Note 1). Compared to other empirically defined sparse representations, the sparse representations learnt in ISTA-Net can encode the neuronal signal more efficiently (Supplementary Fig. 3). Besides the smooth neuropil background, ISTA-Net can also learn to suppress features that are spatially discrete and confined in small volumes, such as punctate cross sections of axons or dendrites (Supplementary Fig. 2, 4; Supplementary Note 1). ISTA-Net processes each frame independently and does not rely on the temporal information. The convolutions in ISTA-Net are effectively 2D. Yet, it achieves a similar performance as other state-of-the-art background suppression networks such as the RB-Net in DeepWonder[29], which requires an additional convolution dimension in the time domain (Supplementary Fig. 4). This reiterates the effectiveness of using the sparse representation.

Besides background suppression, ISTA-Net effectively removes Poisson and Gaussian noise through its convolutional layers by minimizing the mean square error between the output data and the noiseless background-free ground truth data (Supplementary Fig. 2). Its denoising performance is superior to that of DeepCAD, a state-of-the-art denoising algorithm for two-photon calcium imaging (Supplementary Fig. 5). ISTA-Net produces a multi-channel, denoised and background-suppressed video in the sparse representation domain (Extended Data Fig. 2c, 3c, and Supplementary Videos 1-2), with each channel containing the neuronal cell body features. This serves as the basis for the subsequent ConvLSTM2D module to segment individual neurons. Additionally, ISTA-Net outputs a

denoised and background-suppressed video in the spatial domain (Extended Data Fig. 2b, 3b, and Supplementary Videos 1-2) which will be fed to the 1D convolutional layer module for temporal trace extraction.

The second module, a 2D convolutional LSTM (ConvLSTM2D)[38] network, processes the ISTA-Net's output video in sparse representation to segment the neurons and generates frame-by-frame attention maps (Extended Data Fig. 2d, 3d, and Supplementary Videos 1-2) which highlight the active region of each neuron per frame. LSTM is a type of Recurrent Neural Network (RNN)[39] and is commonly implemented to solve the 1D temporal problem of natural language processing[40]. To solve the spatiotemporal sequence prediction problem in calcium recording, we take locality into consideration by using ConvLSTM2D which contains 2D convolutional structures in both the input-to-latent-state and latent-state-to-latent-state transitions[41]. The 2D convolutional layers create latent states from the input of the sparse representation of the recordings, and recurrently process the latent states between neighboring time-steps to learn the autoregressive model of the calcium transient. The sequential prediction of ConvLSTM2D model is further processed by a Cascade Feature Fusion (CFF) layer[42] (Methods), which outputs the segmentation results and the time-series attention maps of individual ROIs. There are two noteworthy features of this module. Firstly, it utilizes a multi-channel sparse representation of the data as input, which contains the pre-processed spatial features of the neurons. This facilitates the ConvLSTM2D to extract the attention maps. This approach significantly improves the segmentation performance over the case where the input to ConvLSTM2D is the same denoised and background-suppressed video but in the spatial domain (single channel) (Supplementary Fig. 6), or where the ISTA-Net is replaced with other non-CS networks such as U-Net[43] and densely-connected Network (DCSRN)[44,45] with the same number of output channels as ISTA-Net (Supplementary Fig. 7). Secondly, the recurrent process provides a powerful mechanism to learn the temporal dynamics of the ROIs and thus the attention maps of each frame, which is the basis for segmentation as well as the subsequent temporal traces extraction. Indeed, compared with conventional 3D CNNs layers, DeepCaImX with ConvLSTM2D has a superior performance in segmentation and traces extraction (Supplementary Fig. 8).

The last module extracts and demixes the neural activity traces of each ROI. This module first performs an overlap integral between the denoised and background-suppressed spatial domain video from ISTA-Net with the time-series attention maps of individual ROIs from ConvLSTM2D to extract their activity traces. As the attention maps report the probability of each pixel being active in each frame, they facilitate signal demixing between the spatially overlapping neurons (Supplementary Fig. 9). Activity traces of all neurons are then fed to a multi-layer 1D CNN, which performs convolution in the temporal domain to further demix the signals, remove residual background and noise, and ensure temporal continuity (Extended Data Fig. 2e, 3e, Supplementary Fig. 9, and Supplementary Videos 1-2). These convolutional layers function similarly to non-negative matrix factorization (NMF) but surpass traditional NMF in demixing overlapping neuronal signals through learning and offering a >10x processing speed, with its advantage growing as the number of neurons within the FOV increases (Supplementary Fig. 10).

Our end-to-end framework significantly streamlines the process, allowing for a fast and efficient neuronal ROI segmentation and activity trace extraction without the need for pre-processing (e.g. spatial and temporal filtering) or post-processing (e.g. merging spatially-adjacent and temporally-correlated ROIs), which may introduce biases. While each module in DeepCaImX has a distinct functionality and is seemingly independent, we train them concurrently as a whole rather than independently or sequentially. This is because some outputs of these modules, such as the sparse representation from ISTA-Net, do not have an accessible ground truth. The loss function is calculated for the denoised and background-suppressed spatial domain video from ISTA-Net, the segmentation from ConvLSTM2D, and the temporal traces from 1D convolutional layers all together against the ground truths in the simulation data. Such a holistic training of the network promotes an overall optimized result.

Segmentation and demixing of neurons with spatial overlaps

We first evaluated DeepCaImX's capability to segment neurons across various sizes and in situations with spatial overlap among neurons or nearby dendrites. Our training data primarily features neurons with diameters set to be 10 to 20 pixels. For experiments with pixel sizes outside our training range, the imaging data can be rescaled to fit these bounds. Our method can reliably and accurately detect neurons within this diameter range (Extended Data Fig. 4). For neurons with a diameter below 10 pixels, the method tends to slightly overestimate the size. The method starts to fail when the neuron diameter approaches below 8 pixels. When the neuron diameter is over 20 pixels, our method underestimates the ROI size and thus only detects the ROIs partially.

We further explored how well our method can demix two neurons with spatial overlap and temporal correlation (Extended Data Fig. 5). Our method accurately distinguishes neurons with weak temporal correlation (Pearson correlation <0.05) and a spatial intersection over union (IoU) up to ~0.346. As the temporal correlation of the two neurons increases, separating neurons becomes more challenging. Yet our model can effectively segment neurons with strong temporal correlation (Pearson correlation 0.6) and substantial spatial overlap (IoU~0.172), achieving a high correlation (~0.94) between the extracted traces and the ground truth. Neurons with identical activities are considered a single unit unless spatially distinct.

In a similar manner, we investigated if our method could segment neurons and decontaminate their activity traces when there are axons or dendrites crossing the neurons in plane or axially (Extended Data Fig. 6). Our algorithm can segment these neurons well and output the activity traces with a high Pearson correlation with the ground truth. This Pearson correlation drops as the contamination increases i.e., when the peak activity intensity ratio between the axons/dendrites and the neurons increases. Nonetheless, even if this ratio is 100%, the neurons can be successfully segmented and the resulting trace extraction has a high correlation with the ground truth (~0.89), demonstrating the strong demixing capability of our algorithms.

As a data driven approach, our methods can be adapted to applications where the ROI size is small or there is a high chance of signal crosstalk between neurons and axons/dendrites. In

those cases, we could specifically tailor the training data so that the model performs well for a given parameter space.

Neuronal segmentation and activity extraction: simulated data

We benchmarked DeepCaImX's segmentation results and the trace extraction performance against existing state-of-the-art methods (STNeuroNet[26], SUNS[27], CITE-On[28], DeepWonder[29], CaImAn[24], Suite2p[20], and methods combining DeepCAD[30], CaImAn[24], and FISSA[23]). All the trainings for the supervised learning methods use the same simulated dataset. We set the intensity of the neuropil such that the data had a similar signal-to-background ratio as that of experimental recordings from a tissue depth of 275 μm (Allen Brain Observatory (ABO) dataset[46]). We tiled each video into smaller sub-videos for processing, and merged the extracted ROIs and traces from all individual sub-videos post-analysis (Extended Data Fig. 7; Methods).

DeepCaImX outperformed other methods in neuron segmentation and activity trace extraction on new simulated testing datasets. Qualitatively, the results closely match the ground truth (Fig. 2a-b). Quantitatively, our method achieves a high precision 0.900 ± 0.015, recall 0.890 ± 0.02 and F1 score 0.892 ± 0.017 (mean±std, $n = 18$ videos) in the segmentation, which are significantly higher than all others (Fig. 2c1–3) [$p<0.05$, one-way Analysis of Variance (ANOVA)]. In other words, our method finds the highest number of accurate ROIs, and among all the ROIs that are found our method has the highest accuracy. Our method also achieves a high Pearson correlation between extracted activity traces and ground truth traces (0.951 ± 0.009) and outperforms others significantly (Fig. 2d) [$p<0.05$, one-way ANOVA], effectively demixing signals even among neurons with large spatial overlap (Fig. 2b). Compared with deep neural networks such as STNeuroNet[26] and SUNS[27] that are specifically designed for segmenting calcium imaging data, DeepCaImX not only has an increased F1 score in ROI segmentation, but also denoises and demixes activity traces within the same network (Fig. 2c3, 2d). While model-based methods such as CaImAn[24] and Suite2p[20] generate comprehensive results in both ROI segmentation and trace extraction, our data-driven method outperforms them in neuropil suppression and ROI segmentation, even when CaImAn is combined with a pre-processing denoising module DeepCAD[30]. Besides its outstanding performance in segmentation and trace extraction (Fig. 2e), DeepCaImX processes the data at 247±8 frames per second, significantly faster than the model-based learning approaches, and competitive with other deep learning segmentation techniques such as SUNS[27].

We investigated why DeepCaImX outperforms other deep neural networks in segmentation, particularly against strong backgrounds. This is not because of hyper-parameter tuning as there are no hyper-parameters in DeepCaImX to tune, unlike other methods such as SUNS[27]. Rather, DeepCaImX could better suppress the background, making the neuronal ROIs more discernible. Successful segmentation requires distinguishing neuronal footprints from the background. Existing deep neural networks, such as SUNS[27] and DeepWonder[29], attempt background suppression through spatial filtering or a separate network, before the actual segmentation. These background suppression procedures operate in spatial domain, making it challenging to distinguish weak neuronal signals from a strong background.

DeepCaImX, however, suppresses the background in the sparse representation, enabling more effective separation of the neuronal signals from background. This is critical to set our method apart from other methods in performance (Supplementary Fig. 11).

Neuronal segmentation and activity extraction: experimental data

We further benchmarked DeepCaImX against others using experimentally recorded calcium imaging datasets from the Allen Brain Observatory (ABO)[46], spanning a wide range of imaging depths from 175 μm to 625 μm (Methods). Ground truth ROIs were manually segmented, and activity traces identified through filtering serve as reference/proxy ground truths (Extended Data Fig. 8; Methods). All supervised learning models, including ours, were trained on the simulated dataset with a wide range of signal-to-background ratios. We then evaluated the performance of each method on the experimental dataset.

Consistent with the simulated dataset results, DeepCaImX outperforms others in the neuronal segmentation and trace extraction on experimental dataset (Fig. 3). It achieves significantly higher precision (0.852 ± 0.031, mean $\pm$ std, $n = 10$videos), recall (0.841 ± 0.045), and F1 score (0.847 ± 0.037) for segmentation, and Pearson correlation (0.938 ± 0.019) between the extracted traces and ground truth traces [$p<0.05$, ANOVA]. As the experimental dataset covers a wider imaging depth range and substantially stronger noise and neuropil background occurs at deeper imaging layers (Supplementary Fig. 12), the F1 score and Pearson correlation show more variance and a drop compared with the simulation test results (Fig. 2) for all methods. Nevertheless, for each sample depth and thus each signal-to-background level, our method outperforms the others (Supplementary Fig. 13).

Leveraging its high-speed operation, we evaluated DeepCaImX on a mesoscopic dataset sample with an FOV of 3×3 mm^2 with 2048×2048 pixels (Extended Data Fig. 9), which is part of a larger 3×5 mm^2 FOV recording from Diesel2p[9]. By segmenting the dataset into 43×43 tiles (with 16 pixels overlap between tiles) and processing each with DeepCaImX, we identified 5175 neurons, at a processing speed of 2.8 frames/sec. This demonstrates our method's scalability and efficiency.

Performance of DeepCaImX in data with different SBRs and SNRs

We investigated how DeepCaImX's performance is influenced by the signal-to-background ratios (SBRs) and signal-to-noise ratios (SNRs) of the dataset (Methods). In the previous section and Fig. 3, we trained a single general model on simulated datasets covering a wide range of SBRs and SNRs. As the SBR or SNR increases in the test dataset, the performance of this general model increases (Fig. 4). We further developed SBR-specific and SNR-specific models of DeepCaImX by separately training different models for specific SBRs or SNRs. We tested these models individually on both simulated test datasets and experimental datasets with correspondingly similar SBRs or SNRs. These specialized models outperformed the general model in both segmentation and trace extraction, particularly for dataset with low SBR or SNR (Fig. 4). Essentially, they could find more ROIs whose SBR or SNR is low (Extended Data Fig. 10), demonstrating the advantage of the data-driven approach.

We further compared the SBR-specific DeepCaImX against other neural-network-based supervised learning methods tailored to the same SBR (Supplementary Fig. 14). The comparison shows that DeepCaImX excelled in both segmentation and trace extraction, reaffirming its superior performance over other neural-network-based methods.

Discussion

We developed DeepCaImX, a multi-task end-to-end model, to segment neuronal ROIs and extract activity traces simultaneously in calcium imaging videos. In comparisons with other state-of-the-art calcium imaging processing algorithms, DeepCaImX excels in performance in both spatial segmentation and temporal activity demixing across simulated and experimental datasets. It also outperforms all but one algorithm in running speed. The superior performance of DeepCaImX makes it an outstanding tool for comprehensive calcium imaging analysis.

DeepCaImX stands out from existing deep learning methods in image and video processing with its multi-task, multi-class and multi-label capabilities. At its core, the ConvLSTM2D module analyzes the spatial correlation between pixels and learns their temporal signal patterns through the recurrent connections. The attention maps it generates could then be processed to obtain segmentation results and extract temporal activity traces (multi-task) of multiple ROIs (multi-class). An individual pixel can appear in the attention maps for different ROIs (multi-label), enabling the segmentation of spatially overlapping neurons. Existing deep learning techniques designed for calcium imaging processing, such as DeepWonder[29], STNeuroNet[26], SUNS[27] and CITE-On[28] are single-input single-output models to focus solely on segmentation, and primarily excel at ROI detection. The temporal trace extraction and demixing are processed after the neural network. Conversely, matrix factorization approaches such as CaImAn[24] and Suite2p[20] simultaneously segment ROIs and demix temporal traces, offering high-quality trace extraction and demixing for those accurately segmented neurons. However, they suffer from higher rates of false positives and negatives in segmentation. DeepCaImX, as a single-input multi-output model, combines the advantages of segmentation neural networks and matrix factorization approaches to achieve superior performance in both segmentation and trace extraction.

Another feature of our method is its robust neuropil suppression capability through a CS-inspired ISTA-Net. In the spatial domain, neuropil appears as a smooth background that mixes with the useful signal. CS leverages sparsity to recover signals from fewer samples than the Nyquist–Shannon sampling theorem requires. Instead of its original application for signal recovery from under-sampled measurements, we employ ISTA-Net to find a sparse representation in which the true signal from neuronal cell bodies is strong and the neuropil is weak. This facilitates an effective separation and suppression of neuropil.

Our approach in segmentation is distinct from existing neural networks. Segmentation in DeepCaImX contains two steps: ISTA-Net to synthesize multiple feature channels of neuronal cell bodies in sparse representation, and ConvLSTM2D to recurrently process these features and generate the temporal dynamic attention maps for each neuron and the static segmentation result. In contrast, other networks typically process the entire video or their

spatiotemporal projections using 2D or 3D convolutional layers[26–29] to identify the neuronal footprints. While generally effective, they may lack efficiency and clarity in parsing the neuronal footprint apart from the background. Conversely, DeepCaImX is tailored to the physical aspects and constraints of calcium imaging, and utilizes a sparse representation of the recordings to analyze the spatiotemporal features of the neurons. By doing so, it can effectively set neurons apart from neuropil background (Supplementary Fig. 11) and improve the subsequent segmentation and temporal activity extraction.

Existing neural networks designed for ROI segmentation extract the temporal activity traces after the segmentation network, using either pixel averaging within neuronal footprints as in STNeuroNet[26] and SUNS[27], or NMF as in CITE-On[28] and DeepWonder[29]. The former could not demix signals in spatially overlapping neurons, and the latter has a slow processing speed. Our pipeline uses the frame-by-frame attention maps and the 1D convolutional layers for efficient demixing, outperforming NMF (Supplementary Fig. 10) and thus other segmentation-focused networks in both speed and effectiveness.

The DeepCaImX is well adapted to datasets with different SBRs and SNRs (Fig. 4). Models that are trained on specific SBRs and SNRs yield better results for data with similar SBRs and SNRs. Ideally, training different models across a range of SBRs or SNRs allows for selecting the most suitable one for a given dataset. However, DeepCaImX's general model, trained on datasets with various SBRs or SNRs, also excels and surpasses other methods (Fig. 3). For datasets with significant SBR variations within a single FOV, the general model may be preferred. Further improving the performance could involve developing a fusion network which synthesizes output from various SBR- or SNR-specific models.

DeepCaImX is user-friendly without any requirements to tune hyper-parameters in the pre-processing or post-processing stages. This method can be adapted to different datasets or applications by adjusting the weights of the loss function of each task (Methods). While its current implementation focuses on segmenting neuronal cell bodies in the mouse cortex, this data-driven algorithm can potentially be extended to other brain regions, species, and subcellular structures. Though developed for two-photon calcium imaging, DeepCaImX is also promising for one-photon calcium imaging where it can distinctly separate the stronger neuropil background from signals in sparse representations. Future research could investigate its efficacy in scenarios like population optogenetics or epileptic states, where neurons and neuropil exhibit highly synchronized activity.

Methods

Simulated datasets of two-photon calcium imaging.

We used NAOMi[36] to synthesize simulated datasets of two-photon calcium imaging. Each dataset (488×488 pixels, 1000 frames) contains 150~350 randomly positioned neurons with diameters of 10~20 pixels in general. This reflects the typical experimental conditions with 0.8~1.5 μm/pixel and 10~15 μm diameters of neuronal cell bodies in mice. Neuronal activities were represented by temporal spikes generated by a Poisson process. The calcium transient kernel of each spike was modeled as exponential functions[23,36] using the rise and decay time of a selected calcium indicator (GCaMP6s or 6f[47], or jGCaMP7b, 7c,

7s or 7f[48]). For each calcium indicator, we also varied the rise and decay time based on their experimental characterization[47,48]. The noise-free fluorescence traces were then generated by convoluting the spikes with a given calcium transient kernel. We then summed all the pixel-wise multiplication results between the neuronal footprints and their corresponding temporal activity traces into a single FOV. The neuropil background was modeled as the summation of dendrite/axon components and an additional 3~5 different background components, each of which is an element-wise product between a Gaussian kernel (100~120-pixel standard deviation with the centroid randomly assigned) and a unique Wiener process. Next, we used a point-spread-function (0.6 excitation numerical aperture) to scan the spatiotemporal data frame-by-frame, and added Gaussian and Poisson noise so the SNR ranged from 3~10. A total of 108 samples with the 6 different types of calcium indicators were used in the training.

Allen Brain Observatory (ABO) experimental dataset and ground truth.

We used the ABO[46] calcium imaging datasets to test the algorithms. The dataset includes 10 videos recorded at 30 Hz from 175 μm, 275 μm, 375 μm, 550 μm and 625 μm deep in the primary visual cortex of 10 mice transfected with the GCaMP6f calcium indicator (sample IDs: 501271265, 501704220, 524691284, 531006860, 603516552, 604145810, 607040613, 669233895, 671162628 and 679353932). Each frame was cropped to 488 × 488 pixels. We selected 5000 consecutive frames for testing. Two human experts manually segmented each recording to create a consensus segmentation ground truth. The manual labeling was performed by inspecting the standard deviation projection of the recordings and then the calcium transients (sharp rise and slow decay) of each ROI found in this projection. We detected additional ROIs by inspecting the recording in small tiles and voted to determine whether each of these new ROIs could be classified as a cell body. To approximate each neuron's ground truth activity trace, we first removed Gaussian and Poisson noise through a bilateral filter frame by frame in the spatial domain; we then estimated each neuron's background component by applying a lowpass filter on the intensity traces of individual pixels within each neuronal footprint and subtracted this background; finally, we calculated the average intensity of all the pixels within the neuronal footprints to obtain the temporal trace for each neuron. For the neurons with spatial overlap with others, we only included the non-overlapped regions in the calculation and proportionally increased its temporal intensity according to the ratio of the overlapping region to the entire region of the individual neuron. We validated this process by comparing the results with ground truth traces in simulated dataset (Extended Data Fig. 8).

Structure of DeepCalmX.

The model consists of 3D CNNs, 2D CNNs, 1D CNNs, nonlinear units, a CFF layer and an Average Pool (Extended Data Fig. 1). The video stack (64×64 pixels, 400 frames) first goes through a 9-phase ISTA-Net in a frame-by-frame manner, where $P = 9$ ISTA-blocks (or phases) with the same architecture are cascaded sequentially, to generate an output where the background and noise are suppressed. In each ISTA-block, each image frame is transformed to a sparse representation domain, where a soft threshold is employed, before being transformed back to the spatial domain. The sparse representation (N channels) resulting from a sparse transformation $\mathcal{F}$ and soft threshold $soft_{\lambda}$, and the spatial domain

output (one channel), both from the last phase in the ISTA-Net, serve respectively as the inputs of ConvLSTM2D and a three-layer 1D CNN. While such a process is in 2D, we used a 3D version of ISTA-Net in the actual implementation as it is more convenient to perform end-to-end training. This 3D network processes the video frame-by-frame in the same way as a 2D network with a kernel size of 1×3×3, where the degree in the temporal dimension is 1. In ConvLSTM2D, we use 2D convolutional layers to replace the fully connected layers typically used together with LSTM to reduce redundant connections and guide the optimization to capture the local information in the spatiotemporal data. This dramatically reduces the learnable parameters and simplifies the training. We use "tanh" and "sigmoid" activations for nonlinearity. Depending on the density of the neurons, we could set up N channels to host N individual neurons in the model. Here, we set N=15, accommodating most imaging settings with up to 15 neurons in a 64×64 pixel image. ConvLSTM2D generates features for attention maps of individual neurons at the corresponding channels in each frame. The features of each channel are then sent to a Cascade Feature Fusion (CFF) layer[42], which is a scene parsing network using a pyramid pooling module and spatially dilated convolution. Here, the CFF layer predicts an attention map for each time frame and perceptually processes the time-series attention maps into a multi-channel ROI projection, with each channel containing the segmentation result of an ROI. The CFF layer further utilizes a morphological operation, called opening[49], to remove the small discrete area that are separated from the main ROI, and to obtain the final segmentation results. Using the time-series attention maps, and the multi-channel ROI projections, we could generate the time-series attention maps for each ROI channel. If the number of neurons found is smaller than the total channel number, the remaining channels are left empty. The overlap integral between the time-series attention maps for each ROI and the spatial domain video from ISTA-Net produces the activity traces for individual neurons. Finally, we use a 1D three-layer CNN, with "relu" activations after each of the first two layers, to demix neuronal activity traces from residual contamination in the FOV. The convolution is conducted in the time domain.

To accommodate videos with different pixel resolutions and counts, we first scale the video so its spatial resolution falls in 0.8~1.5 μm/pixel, and thus most neurons have a cell body of 10~20 pixels in diameter. We then tile the entire video into different sub-stacks in both the spatial and temporal directions, each being 64×64 pixels and 400 frames. Zero padding will be applied if the sub-stacks have fewer than 64×64 pixels and 400 frames. Spatial overlap occurs between neighboring sub-stacks. Each sub-stack is processed by DeepCaImX, yielding ROI segmentation and activity traces. The results from all 3D sub-stacks are then merged together. During merging, neurons in the overlapping regions of neighboring sub-stacks are matched if their activity traces have a Pearson correlation above 0.95. Each matched neuron pair is merged into a single neuron, with combined spatial footprints and weighted sum of temporal activity traces, where the weight is based on the area of individual segmentations and the union. Neurons detected in different time slots with an IoU above 0.9 for their spatial footprints are considered the same neuron, with combined footprints and concatenated activity traces. If a neuron is not detected in a time slot for the entire 400 frames, its activity trace is set to 0 for that particular time slot. The results of the entire video can be obtained after this spatial merging and temporal concatenation process.

In the training dataset, we separated each original 488(pixel)×488(pixel)×400(frame) video into 81 64×64×400 3D sub-stacks with 11 pixels overlap in each spatial dimension. We trained DeepCaImX for 20 hours with a batch size of 2 and a learning rate of 1×10^{-4}.

In the simulated dataset used for testing, each video has a size of 488 (pixel) × 488 (pixel) × 1,000 (frame). We temporally tiled the video into three time slots, each having 400 frames. The temporal overlap is 0 frames, and the last 200 frames of the last time slot are set to be 0. For each time slot, we performed spatial tiling with the same tiling setting as those in the training dataset.

In the ABO experimental dataset used for testing, each video has a size of 488(pixel) ×488(pixel)×5000(frame). We temporally tiled the video into 13 timeslots, each having 400 frames, and kept the spatial tiling setting the same as those in the training dataset. In total, 1053 sub-videos were created for each dataset.

In the Diesel2p mesoscopic recordings used for testing, the sample has a FOV of 3×3 mm^2 and a size of 2048(pixel)×2048(pixel)×1500 (frame), which is a subset of a 3×5 mm^2 FOV data. We spatially tiled the dataset into 43×43 tiles (with 16 pixels overlap between tiles). Temporally, we tiled the video into 4 timeslots, each having 400 frames. In total, 7396 sub-videos were created.

Loss function of DeepCaImX.

The loss function of DeepCaImX is the weighted sum of three parts: (1) the loss function of ISTA-Net, (2) the dice coefficient[50] of ROI prediction, and (3) the Pearson correlation between the extraction results of all traces and the ground truth. Mathematically, the loss function is written as:

$$Loss1 = \left\| \boldsymbol{x}^{(P)} - \boldsymbol{x} \right\|_2^2 + \gamma \left(\sum_{k=1}^{P} \left\| \mathscr{F}^{-1^{(k)}} \left(\mathscr{F}^{(k)}(\boldsymbol{x}) \right) - \boldsymbol{x} \right\|_2^2 \right) \tag{1}$$

$$Loss2 = \left(1 - \frac{2\left| seg^{Pred} \cap seg^{GT} \right|}{\left| seg^{Pred} \right| + \left| seg^{GT} \right|} \right) \tag{2}$$

$$Loss3 = 1 - \frac{\sum_i trace_i^{Pred} trace_i^{GT} - \sum_i trace_i^{Pred} \sum_i trace_i^{GT}}{\sqrt{\sum_i (trace_i^{Pred})^2 - \left(\sum_i trace_i^{Pred}\right)^2} \sqrt{\sum_i (trace_i^{GT})^2 - \left(\sum_i trace_i^{GT}\right)^2}} \tag{3}$$

$$Loss = Loss1 + \lambda_1 Loss2 + \lambda_2 Loss3$$

where $\boldsymbol{x}^{(P)}$ represents the reconstructed spatiotemporal recording after the P^{th} phase of the ISTA-Net, $\boldsymbol{x}$ represents the background-free and noise-free ground truth, $\mathscr{F}$ is the learnable sparse transformation supported by ISTA-blocks, $\mathscr{F}^{-1}$ is the learnable backward transform from sparse representation to the original spatial domain, k is the phase index of the ISTA-block, P is the number of the ISTA-block phase, and γ is a weight of 0.1 in the loss function of ISTA-Net. seg^{Pred} and seg^{GT} represent the predicted segmentation result and the corresponding ground truth, respectively. $\text{trace}^{\text{Pred}}$ and trace^{GT} represent the predicted activity trace and the corresponding ground truth, respectively i in Loss3 means the i-th entry of the traces. λ_1 and λ_2 are the relative weights of the loss function (2) and (3) versus (1). λ_1 and λ_2 are both set to be 10 (see further discussion in "Training, setting and modifying DeepCaImX").

Evaluation metrics of ROI segmentation and temporal activity trace extraction.

We evaluated all segmentation methods by comparing their results with ground truth labels. The metrics of evaluation are recall, precision and F1 score, which are defined as follows:

$$Precision = \frac{N_{TP}}{N_{detected}},$$

$$Recall = \frac{N_{TP}}{N_{GT}},$$

$$F1 = \frac{2}{Precision^{-1} + Recall^{-1}},$$

where N_{TP} is the number of true positive (TP) predictions, N_{GT} is the number of ground truth (GT) ROIs, and $N_{detected}$ is the number of neurons detected by the method. To determine if a predicted neuron belongs to the ground truth, we use the IoU metric:

$$IoU(GT, Prediction) = \frac{\left|seg^{GT} \cap seg^{Pred}\right|}{\left|seg^{GT} \cup seg^{Pred}\right|}.$$

If the IoU value is greater than 0.5, we regard the prediction of the neuron to be accurate. If there is more than one neuron whose IoU is greater than 0.5 for a specific ground truth neuron, we will select the one with higher IoU and count that as the true positive.

The evaluation of extracted traces is based on Pearson correlation:

$$r\left(trace^{Pred}, trace^{GT}\right) = \frac{\sum_i trace_i^{Pred} trace_i^{GT} - \sum_i trace_i^{Pred} \sum_i trace_i^{GT}}{\sqrt{\sum_i (trace_i^{Pred})^2 - \left(\sum_i trace_i^{Pred}\right)^2} \sqrt{\sum_i (trace_i^{GT})^2 - \left(\sum_i trace_i^{GT}\right)^2}},$$

where i is the i-th trace entry.

Training, setting and modifying DeepCaImX.

Training DeepCaImX involves two steps: dataset tiling and neural network training. Each raw calcium imaging recording is first tiled into sub-videos both spatially and temporally based on the video size that DeepCaImX is designed for while considering the spatial and temporal overlap for each sub-video. In this paper, the input video size for DeepCaImX is 64×64×400, smaller than the simulated raw video (488×488×400). With 11-pixel spatial overlap along each dimension, 81 tiles with 64×64 pixels are generated. The total epochs and batch size for training DeepCaImX can be adjusted based on available resources.

Hyper-parameters in the loss function could extensively influence the performance of DeepCaImX. The weights λ_1 and λ_2 control the contribution of ROI detection and trace extraction respectively to the overall loss function. We make the values of each of these losses match the loss of ISTA-Net when the optimization is stable, ensuring equal contribution to the multi-task optimization (Supplementary Fig. 15). In this paper, both λ_1 and λ_2 are set to be 10, as the loss of ISTA-Net converges to a range of 7~13. For different types of datasets, the settings of λ_1 and λ_2 may be different. For datasets with high noise, we can increase the relative contribution of the denoising module (i.e. ISTA-Net) by decreasing λ_1 and λ_2 accordingly. For applications requiring very precise ROI predictions, we could increase the weight of the ROI detection loss (i.e. λ_1).

Besides the weights, users can add constraints to the loss function to customize DeepCaImX for specific applications. For example, an ROI area constraint can be added to $Loss2$ when the requirements of ROI area is critical. Additionally, $L1$-norm regularization can be used to enforce sparsity of extracted traces.

General, SBR-specific or SNR-specific DeepCaImX models.

Calcium imaging data varies in SBR and SNR. The signal relates to the brightness and expression level of calcium indicators, and the change in fluorescence due to action potentials. The background relates to the neuropil intensity, influenced by the fluorescence labeling density, imaging depth and animal preparation procedures. The noise can include shot noise (Poisson noise), amplification noise and read noise, and can be modeled as a mixture of Poisson and Gaussian noise. From the raw recordings, we defined the signal as the peak value of each temporal trace, the background as the average neuropil intensity (obtained by lowpass filtering the raw experimental recordings temporally), and the noise as the standard deviation of the difference between the raw signal and the noise-free ground-truth (obtained by applying a 2D bilateral filter to the raw experimental recordings spatially). For each data sample, the SBRs and SNRs for individual neuron are first calculated, and their values are then averaged to represent the SBR and SNR of the entire dataset.

By adjusting neuropil background or noise level, we can create simulated datasets with different SBRs and SNRs. DeepCaImX models specific to SBR or SNR are trained using corresponding datasets. The general model is trained using datasets across a broad SBR and SNR range.

To choose the appropriate DeepCaImX model, users could first estimate the SBRs or SNRs of neurons in the FOV, generate a histogram and determine the average SBRs and SNRs. The suitable model can be chosen based on these values. As a reference for SBR-specific models, users can compare the SBR histogram with models in Supplementary Fig. 12, and select the most similar one. If the SBR variations are large, the general model may be preferred.

Hardware and processing speed calculation.

The processing for all methods in this paper was done on a workstation with an Inter(R) Xeon(R) E5-2667 v3 @ 3.20GHz CPU and a NVIDIA Quadro RTX 8000 48 GB GPU. We included only the runtime calculation of the algorithm without data loading and writing when creating the processing speed profile for each method.

Methods used for comparison against DeepCaImX.

Nine methods are used to benchmark the performance of DeepCaImX: DeepWonder[29], STNeuroNet[26], SUNS[27], CITE-On[28], CaImAn[24], Suite2p[20], and three methods combining DeepCAD[30], FISSA[23] and CaImAn[24]. All training-required methods were trained for 20 hours with 488×488 pixel FOVs. Hyper-parameters of each method were tuned to fit each dataset's properties, including pixel size, frame rate, and calcium indicator type. Below, we describe each method's mechanisms and parameters.

DeepWonder[29] cascades two independently trained 3D CNNs, RB-Net and NS-Net, to perform noise and background subtraction and segmentation respectively. We used the code found at https://github.com/yuanlong-o/Deep_widefield_cal_inferece. The input and ground truth for RB-Net are the raw recordings and the denoised, background-removed videos, respectively. The input and ground truth for NS-Net are the output of RB-Net and the ROI segmentation targets, respectively. Other parameters were set according to the properties of our training and testing datasets (i.e. pixel size, frame rate, calcium indicator).

STNeuroNet[26] applies a 3D CNN with preprocessing and postprocessing to segment the calcium imaging data. We used the code found at https://github.com/soltanianzadeh/STNeuroNet. In the preprocessing stage, we set the size of the gaussian kernel filter to be 20 pixels, and the rise and decay time of the neuronal activity based on the calcium indicator. In the postprocessing stage, we set the minimum area of ROIs of 100 pixels and probability threshold of 0.9. The block size for training is 488×488×100.

SUNS[27] operates fast by using a shallow U-Net but requires hyper-parameter tuning in pre- and post-processing. We used the code found at https://github.com/YijunBao/Shallow-UNet-Neuron-Segmentation_SUNS. We set the SNR threshold to be 3, the minimum ROI area to be 100 pixels, probability threshold to be 0.7 and threshold for maximum center of mass (COM) distance to be 4 μm. The bandwidth of the low pass filter in the preprocessing stage was tuned independently to remove the noise and background for every video.

CITE-On[28] supports very fast and accurate segmentation based on 2D projections of calcium imaging recordings. We used the code at https://gitlab.iit.it/fellin-public/cite-on. We set the

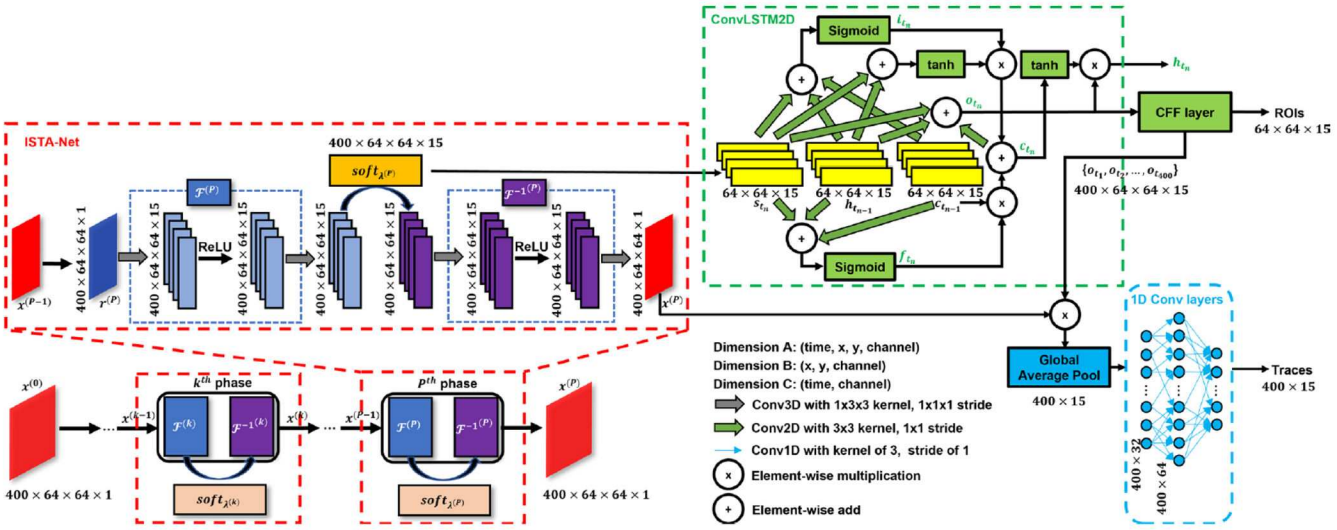
upscaling factor to be 2, tiles per side as 10 with an overlap between tiles of 20%, and a batch size of 16 in the extractor. The other settings followed the properties of our dataset.

CaImAn[24] applies a constrained non-negative matrix factorization (CNMF) algorithm[21] to detect ROIs and extract their temporal activity traces. In CNMF[21], the recording is factorized into the product of two matrices which are related to the spatial footprints and temporal traces of each neuronal ROI and the background components. Regularization is used to promote sparsity of the two matrices. The weights of the regularization term control the degree of sparsity and edge sharpness of each ROI's spatial contour. A CNN classifier filters out ROIs that are less likely to be neurons. We used the CaImAn batch method from the code found at https://github.com/flatironinstitute/CaImAn. Each patch has 64×64 pixels with 11 pixels of overlap between patches. We set the number of components per patch to 20, the spatial correlation threshold to ~0.8, the minimum SNR to ~4, and the upper and lower threshold for the CNN classifier to 0.8~0.9 and 0.1~0.3 respectively, which varied between test samples. Other settings were tuned to fit the properties of our dataset similarly to previous methods. An optional preprocessing step to subtract the global background of the data before CaImAn may enhance its overall performance. A post-processing step may be required when the predicted traces exhibit an unexpectedly prolonged decay time in the calcium transient, which was related to the unsatisfactory background estimate and/or the imbalance between the spatial matrix and temporal matrix in the matrix factorization process. To address this, we normalized the individual predicted traces on a timestep-by-timestep basis by a ratio between the average value of the predicted spatial matrix and the average value of the denoised and background-suppressed raw recordings within the matching ROIs.

Suite2p[20] is based on a matrix factorization algorithm with fewer constraints than CNMF[21]. We used the code found at https://github.com/MouseLand/suite2p. We binarized the real-valued mask output with a threshold set to 0.3 times the maximum value of the mask to obtain the ROI contours. We then used the default classifier (with the diameter of neurons set to be 12.5 pixels) before temporal signal extraction. The same pre-processing step and post-processing step as the CaImAn described above may be used to enhance its overall performance.

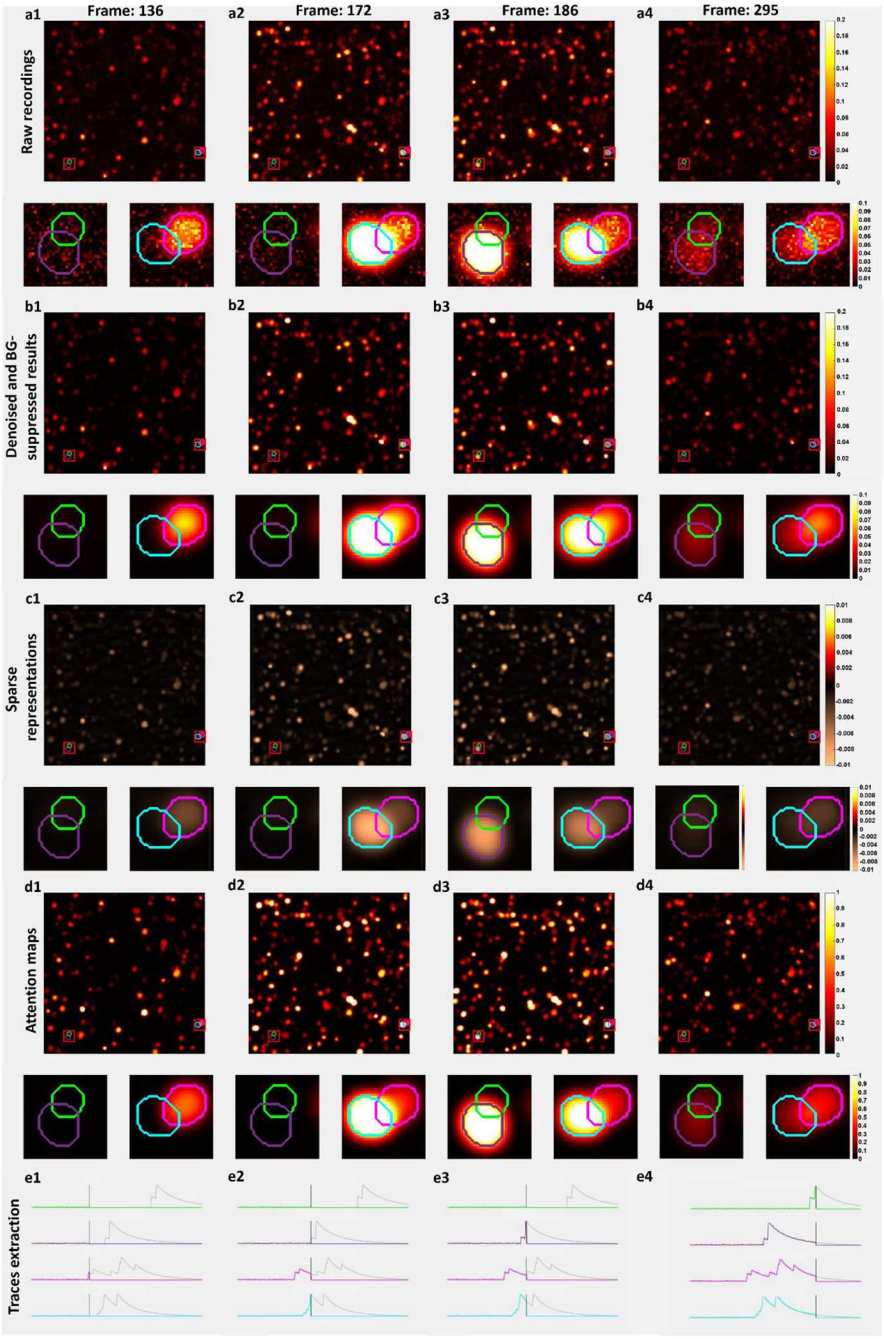
DeepCAD[30] and FISSA[23] are two assistive tools for spatial denoising and temporal demixing with denoising, respectively. DeepCAD uses a self-supervised learning scheme to perform spatial denoising. FISSA uses NMF to decontaminate the neuropil background and extract the temporal signals from ROIs with known spatial contours. We combined these tools with CaImAn to enhance its performance. We used the code found at https://github.com/cabooster/DeepCAD for DeepCAD, and the code found at https://github.com/rochefort-lab/fissa for FISSA. In DeepCAD-CaImAn-assisted FISSA, DeepCAD first denoises the raw data, CaImAn then segments the ROIs, and FISSA extracts each ROI's temporal signals. In DeepCAD-assisted CaImAn, DeepCAD first denoises the raw data, and CaImAn then segments the ROIs and extracts their temporal signals. In CaImAn-based FISSA, CaImAn first segments the ROIs, and FISSA extracts each ROI's temporal signals. The settings in CaImAn were tuned as described previously.

Extended Data



Extended Data Fig. 1 |. Model architecture of DeepCaImX.

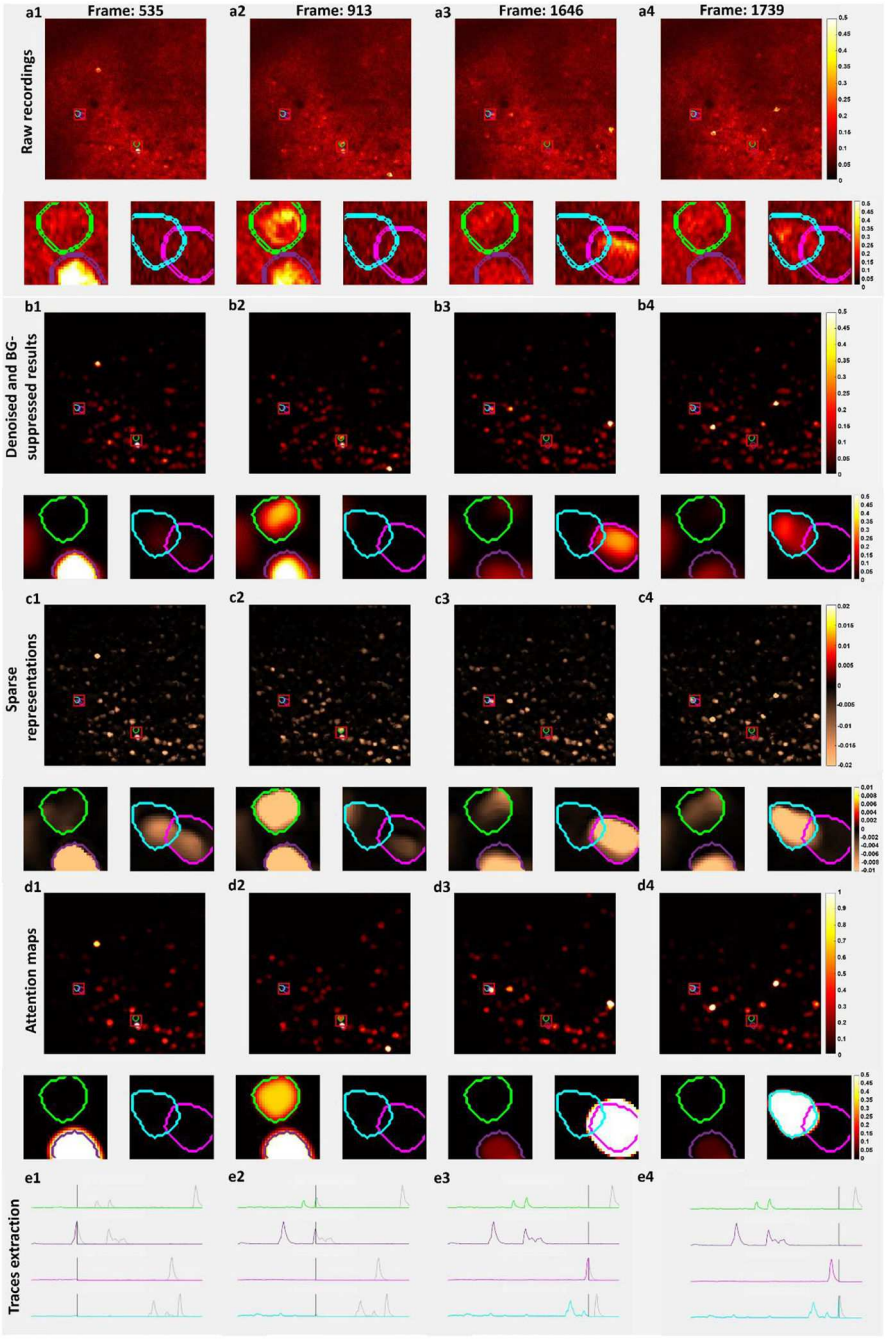
DeepCaImX is composed of three networks: a 3D ISTA-Net, a ConvLSTM2D (2D convolutional LSTM) network, and a 1D convolutional layers. The 3D ISTA-Net is used to suppress the background and suppress the noise of the video stack. This 3D network processes the video frame-by-frame in a way of a 2D network, as the kernel size is $1 \times 3 \times 3$, where the degree of the temporal dimension is 1. It outputs the denoised and background-suppressed video in both sparsity domain (with 15 channels) and spatial domain. The former is fed to the ConvLSTM2D, which outputs the segmentation results through a cascade feature fusion (CFF) layer, and the attention maps. The latter, together with the attention maps, were fed to the 1D convolutional layers, which further demix the temporal signals of the segmented ROIs. The 3D convolutional layers contain a 3x3x1 kernel with stride of 1x1x1. The 2D convolutional layers contain a 2x2 kernel with the same stride. The 1D convolutional layers contain a kernel of 3 and the stride of 1. In ISTA-Net, there is a total of P phases/blocks, with k being the phase/block index. $\mathscr{F}^{(k)}$ means the trainable transformation from the original domain to sparse representation domain, and $\mathscr{F}^{-1^{(k)}}$ represents the inverse transformation. $soft_{\lambda}$ is the soft threshold, with λ being the threshold set to be 0.01.



Extended Data Fig. 2 |. Raw recordings, denoised and background-suppressed video, sparse representation, attention maps and extracted activity traces of represented neurons of a simulated sample.

a1-4, Raw recordings at the frame 136, 172, 186, and 295. Inset shows the exemplary segmented neurons. **b1-4,** Denoised and background-suppressed results at the frame 136, 172, 186, and 295. **c1-4,** Sparse representation results created by ISTA-Net at the frame 136, 172, 186, and 295; **d1-4,** Attention map results created by ConvLSTM2D (2D convolutional LSTM) at the frame 136, 172, 186, and 295. **e1-4,** Extracted traces for the exemplary segmented neurons with cursors at the frame 136, 172, 186, and 295. For an individual

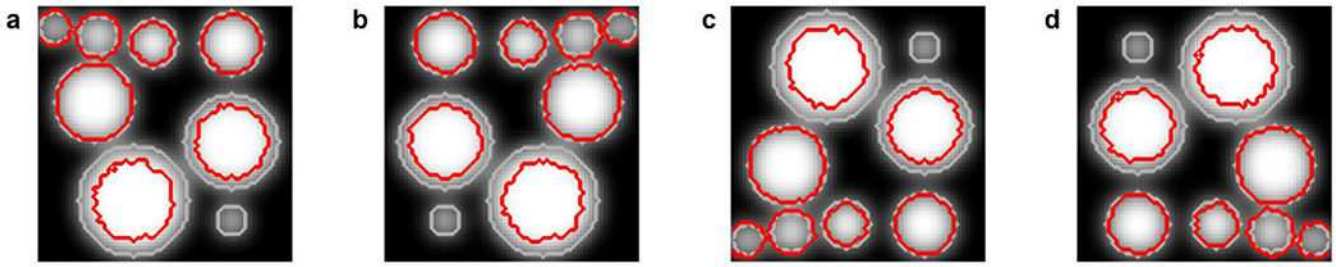
neuron, we used the same color to outline the boundary of its footprint in the different frames and extracted traces. This simulated dataset is the same as that in Fig. 2 in the main manuscript.



Extended Data Fig. 3 |. Raw recordings, denoised and background-suppressed video, sparse representation, attention maps, and extracted activity traces of represented neurons of the experimental sample (ABO 524691284).

a1-4, Raw recordings at the frame 535, 913, 1646 and 1739. Inset shows the exemplary segmented neurons. **b1-4,** Denoised and background-suppressed results at the frame 535,

913, 1646 and 1739. **c1-4,** Sparse representation results created by ISTA-Net at the frame 535, 913, 1646 and 1739; **d1-4,** Attention map results created by ConvLSTM2D (2D convolutional LSTM) at the frame 535, 913, 1646 and 1739. **e1-4,** Extracted traces for the exemplary segmented neurons with cursors at the frame 535, 913, 1646 and 1739. For an individual neuron, we used the same color to outline the boundary of its footprint in the different frames and extracted traces. This simulated dataset is the same as that in Fig. 3 in the main manuscript.



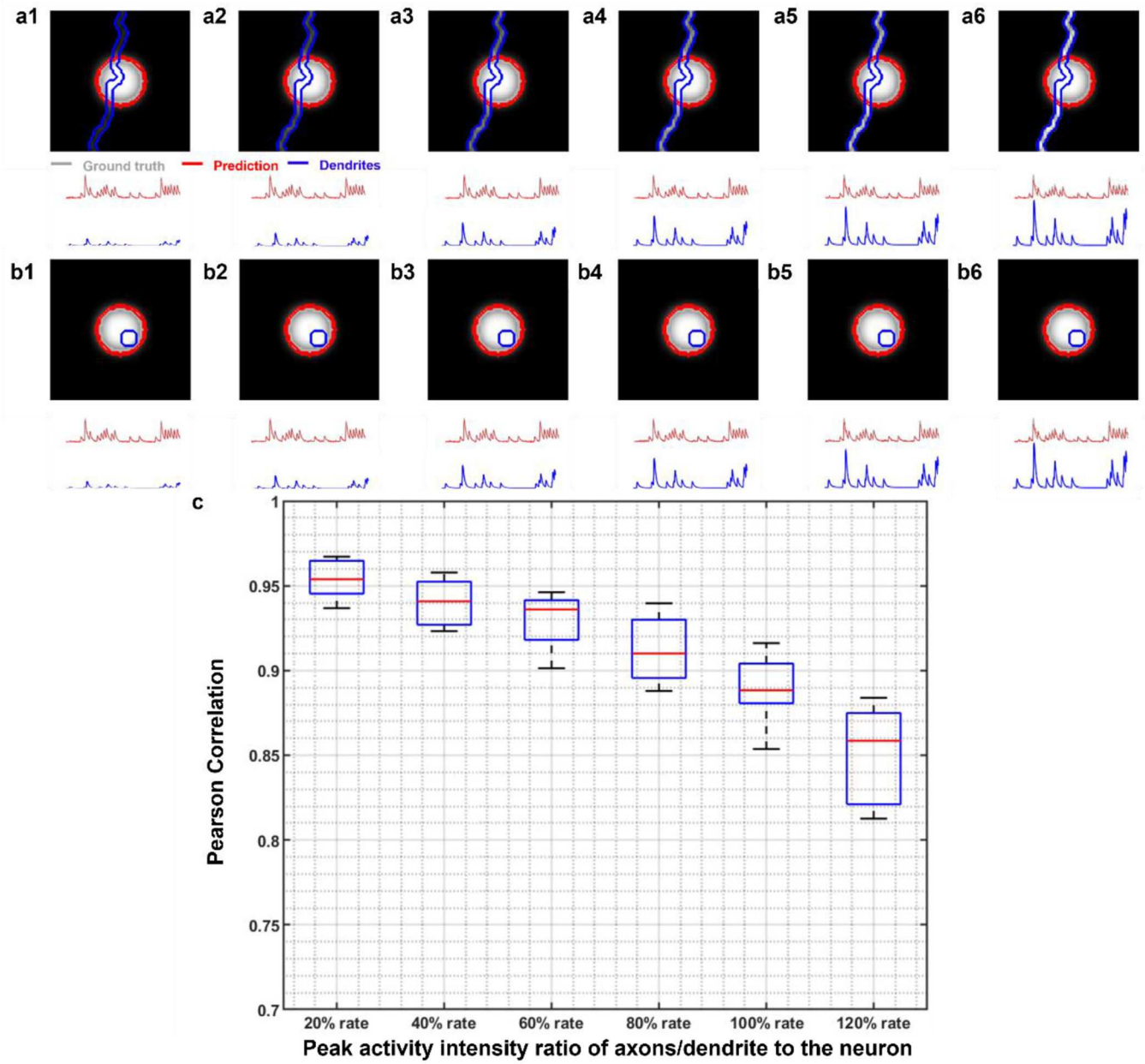
Extended Data Fig. 4 |. Performance of DeepCaImX in detecting neuronal body with different sizes.

When DeepCaImX is trained with a dataset where most of the neuronal body has diameters of 10-20 pixels, it could reliably and accurately detect neurons with a diameter of 8~20 pixels. The predicted segmentation tends to have a larger or smaller boundary than the ground truth when the neuron diameter is <10 pixels or > 20 pixels respectively. **a-d** shows the segmentation results (red line) versus the ground truth (gray line) for neurons with 8 different diameters with 7, 8, 10, 12, 16, 20, 24, and 28 pixels with the same activity trace but in different locations within a field of view.

Extended Data Fig. 5 |. Performance of DeepCaImX in demixing neurons with spatial overlaps and temporal correlations.

We varied the centroid separation of two neurons and studied how well DeepCaImX could distinguish the neurons and demix their activity traces. In **a1-a6**, the centroid separation of the two neurons is 4, 8, 12, 16, 20, and 24 pixels, and each neuron has a diameter of 20 pixels with the same neural activity. The prediction results and ground truth are in red lines and gray lines respectively. In **b1-b6**, **c1-c6, d1-d6, e1-e6**, and **f1-f6**, all the settings are the same except that the temporal correlation between the activity of the two neurons is 0.8, 0.6,

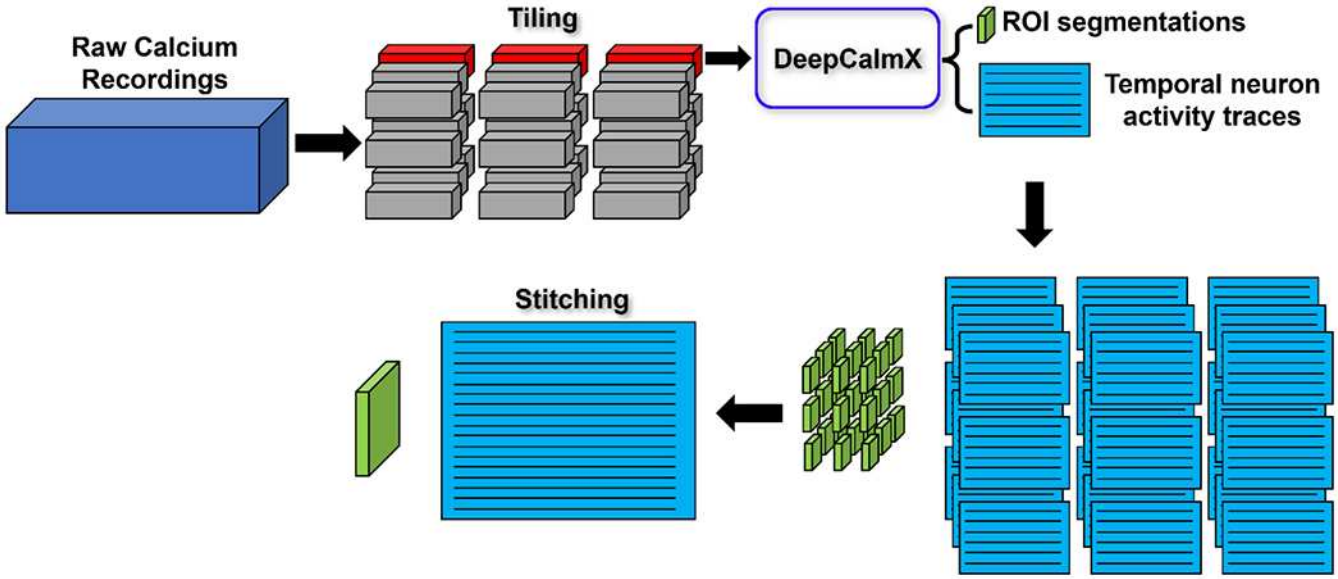
0.4, 0.2, and 0.05 respectively. **g1**. The maximum Intersection over Union (IoU) of the two spatially overlapping neurons that DeepCaImX can segment, versus the Pearson correlation between the ground truth activity traces of the two neurons. **g2**. The Pearson correlation of the extracted activity traces of the two spatially overlapping neurons against their ground truth traces, versus the Pearson correlation between the ground truth activity traces of the two neurons. Data are presented as mean values +/− standard deviation (error bar). IoU is defined as the ratio between the ratio of the intersection area of two neurons to the union region area of two neurons. For **g1-g2**, we created 50 pairs of neuronal footprints by the NAOMi algorithm instead of simply using round-shaped simulated neurons for the study. When the two neurons have high spatial overlap, DeepCaImX may still be able to predict two separated neurons, though their IoU with the corresponding ground truth could be less than 0.5, and thus not considered to be a correct segmentation.



Extended Data Fig. 6 |. Segmentation and demixing the soma activities from nearby axons/dendrites.

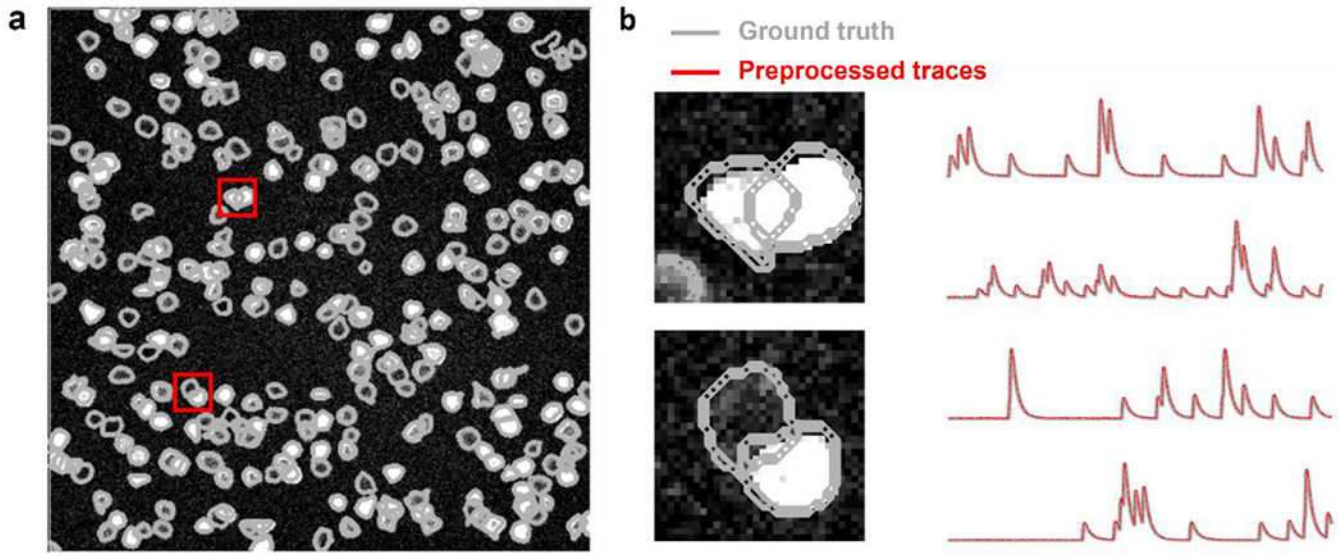
a1-a6, ROI (soma) detection and temporal traces demixing performance of the neuronal soma from the axons/dendrites, for different peak activity intensity ratio (20%, 40%, 60%, 80%, 100% and 120%) between the axons/dendrites and neuronal soma. The axons/dendrites cross the soma in plane. **b1-b6**, Same as **a1-a6**, but with the axons/dendrites cross the soma axially. Here, the cross-section of the axons/dendrites locate inside the neuronal soma to be segmented. **c**, Pearson correlation between the extracted temporal traces of the soma versus the ground truth trace, for different peak activity intensity ratio between the axons/dendrites and the neurons based on 20 simulated samples. Box plot: center bars

(red), medians; box edges, first and third quartiles, respectively; whiskers, minimum and maximum.

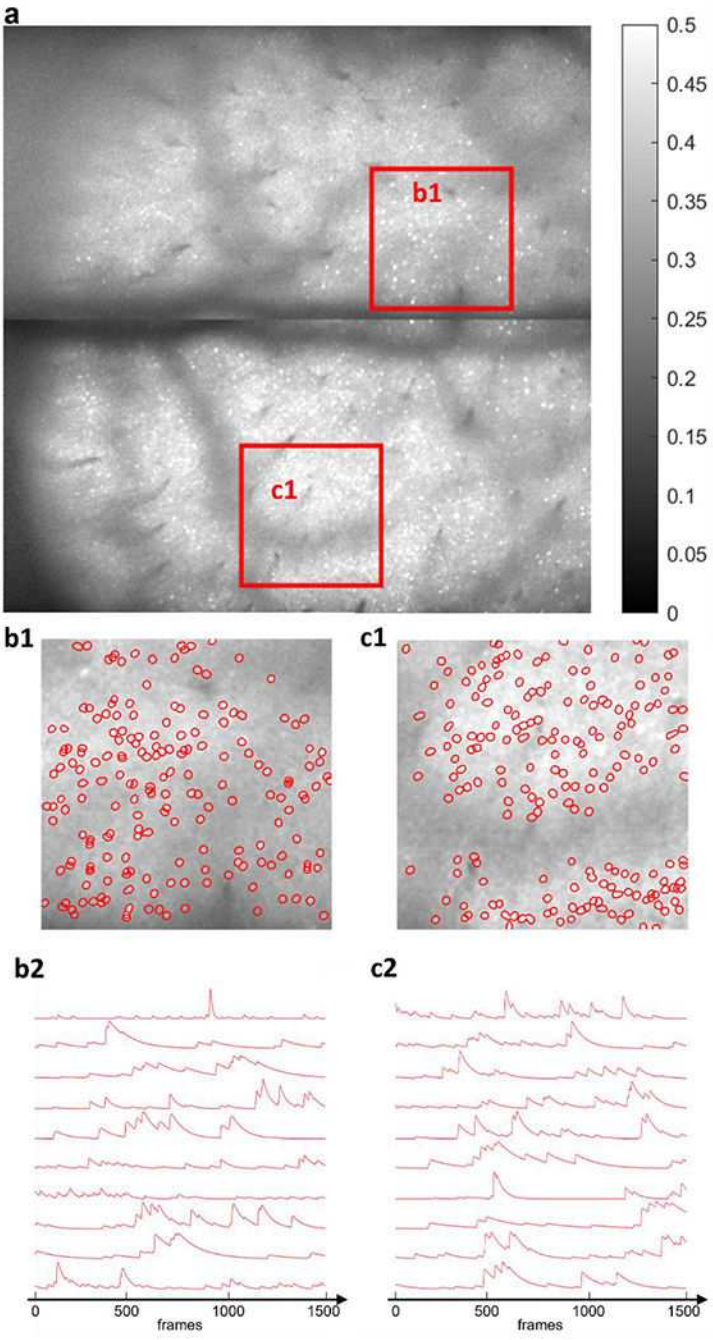


Extended Data Fig. 7 |. Data processing pipeline.

The calcium recording is first tiled into sub-stack of videos with smaller size in space (x, y) and time (t). Each sub-stack of videos is fed to DeepCaImX for spatial segmentation and temporal trace extraction. The results are then stitched together in spatial and temporal dimensions. In our example, the size of experimental raw calcium recording is 488x488x5000 in 3D (x, y, t). This is tiled into 9x9x13 sub-stacks where each sub-stack is in a size of 64x64x400 with an overlapping of 11 pixels in each lateral dimension of the stack, and no overlapping in temporal dimension. In the merging process, for every pair of spatially neighboring sub-stacks in the same time slot, we first match the neurons that are found in both sub-stacks in their overlapping regions if the Pearson correlation of their activity traces is larger than a threshold set at 0.95. We then merge each matched neuron pair into a single neuron whose spatial footprint is set to be the union of the segmentation results from each sub-stack. We generate their temporal activity traces as the weighted sum of the individual traces extracted in each sub-stack, where the weight is based on the area of segmentation results and the union. For any two neurons detected from different time slots, if the intersection over union (IoU) of their spatial footprints is larger than 0.9, we consider that they are the same neuron, and we set the footprint of this neuron to be the union of the segmentation results from each time slot. We then concatenate the activity traces of this neuron in these sub-stacks. Otherwise, we consider they are different neurons. If a neuron is not detected in a specific time slot for the entire 400 frames, we set its activity trace to be 0 for that particular time slot as the neuron is inactive in that time slot. The results of the entire video can be obtained after this spatial merging and temporal concatenation process.



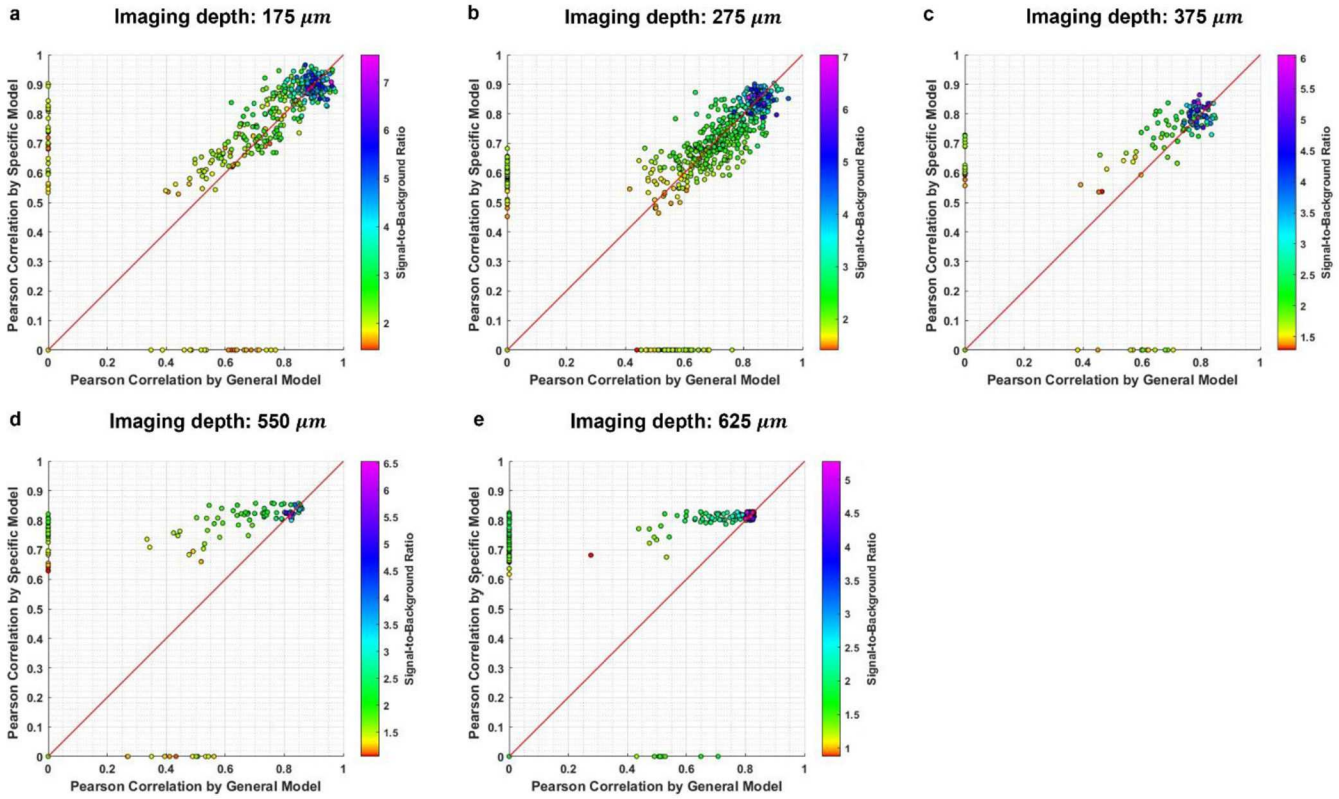
Extended Data Fig. 8 |. Validation of the ground truth temporal traces generation.
We use the process described in Methods to generate the temporal activity traces as proxies of ground truth in the experiment dataset. Here we validated this process in simulation datasets. **a,** The correlation image of a simulated sample. The correlation image is the averaged temporal correlation between pixels and their four immediate neighbors. The boundary of each ROI is annotated with gray contour lines. **b,** Zoom-in region of the red box region in **a** and comparison between the simulated ground truth temporal traces (gray) and the extracted proxies (red). The Pearson correlation between the simulated ground truth temporal traces and the extracted proxies is 0.973±0.018 (mean±std, $n = 18$ videos).



Extended Data Fig. 9 |. ROIs detection and activity traces extraction of Diesel2p mesoscopic two-photon imaging recordings.

a, maximum intensity projection of a recording with a dimension of 3x3 mm^2 and a total pixel count of 2048x2048, which is a subset of the imaging data of a 3x5 mm^2 field of view recorded from Diesel2p, combining the sample of 3a_Ch01 and 3a_Ch02 [Ref. [9], Yu, CH., Stirman, J.N., Yu, Y. et al. Diesel2p mesoscope with dual independent scan engines for flexible capture of dynamics in distributed neural circuitry. Nat Communications 12, 6639 (2021).]. **b1**, The maximum intensity projection of the 1st zoom-in view in (a), with

predicted ROIs overlaid. **c1**, The maximum intensity projection of the 2nd zoom-in view in (a), with predicted ROIs overlaid. **b2**, Extracted temporal activity traces from 10 randomly selected ROIs in the 1st zoom-in view in (a). **c2**, Extracted temporal activity traces from 10 randomly selected ROIs in the 2nd zoom-in view in (a).



Extended Data Fig. 10 |. Pearson-correlation performance of 10 ABO experimental samples via general model vs. SBR-specific models of DeepCaImX.

a-e, correspond to samples recorded in depths of 175 μm, 275 μm, 375 μm, 550 μm, and 625 μm respectively. For each individual segmented neuron in each model, we calculated the Pearson-correlation between the extracted temporal activity traces versus the reference ground truth traces. Each individual segmented neuron is plotted as a point in each plot, with the color indicating the signal-to-background ratio (SBR). The neurons found by the SBR-specific models but not the general model are assigned a Pearson correlation value of 0 for the general model. The neurons found by the general model but not the SBR-specific models are assigned a Pearson correlation value of 0 for the SBR-specific models.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work is supported by National Institute of Neurological Disorders and Stroke (R01NS118289, R01NS133924, W.Y.), National Eye Institute (R01NS118289, W.Y.), National Science Foundation (CARRER 1847141, W.Y.), and Burroughs Wellcome Fund (Career Award at the Scientific Interface 1015761, W.Y.).

Data availability

The ABO dataset (experimental dataset) can be found in

https://github.com/AllenInstitute/AllenSDK/wiki/Use-the-Allen-Brain-Observatory-%E2%80%93-Visual-Coding-on-AWS.

The dataset used in the simulation and training could be found in https://zenodo.org/records/12650420[51].

Reference

1. Denk W, Strickler JH & Webb WW Two-photon laser scanning fluorescence microscopy. Science 248, 73–76 (1990).
2. Grienberger C & Konnerth A Imaging calcium in neurons. Neuron 73, 862–885 (2012).
3. Stosiek C, Garaschuk O, Holthoff K & Konnerth A In vivo two-photon calcium imaging of neuronal networks. Proceedings of the National Academy of Sciences 100, 7319–7324 (2003).
4. Svoboda K & Yasuda R Principles of two-photon excitation microscopy and its applications to neuroscience. Neuron 50, 823–839 (2006).
5. Yuste R & Denk W Dendritic spines as basic functional units of neuronal integration. Nature 375, 682–684 (1995).
6. Beaulieu DR, Davison IG, Kılıç K, Bifano TG & Mertz J Simultaneous multiplane imaging with reverberation two-photon microscopy. Nature methods 17, 283–286 (2020).
7. Wu J et al. Kilohertz two-photon fluorescence microscopy imaging of neural activity in vivo. Nature methods 17, 287–290 (2020).
8. Demas J et al. High-speed, cortex-wide volumetric recording of neuroactivity at cellular resolution using light beads microscopy. Nature Methods 18, 1103–1111 (2021).
9. Yu C-H, Stirman JN, Yu Y, Hira R & Smith SL Diesel2p mesoscope with dual independent scan engines for flexible capture of dynamics in distributed neural circuitry. Nature communications 12, 6639 (2021).
10. Han S, Yang W & Yuste R Two-color volumetric imaging of neuronal activity of cortical columns. Cell reports 27, 2229–2240. e2224 (2019).
11. Prevedel R et al. Fast volumetric calcium imaging across multiple cortical layers using sculpted light. Nature methods 13, 1021–1028 (2016).
12. Weisenburger S. et al. Volumetric Ca2+ imaging in the mouse brain using hybrid multiplexed sculpted light microscopy. Cell 177, 1050–1066. e1014 (2019).
13. Ji N, Freeman J & Smith SL Technologies for imaging neural activity in large volumes. Nature neuroscience 19, 1154–1164 (2016).
14. Yang W & Yuste R In vivo imaging of neural activity. Nature methods 14, 349–359 (2017).
15. Thevenaz P, Ruttimann UE & Unser M A pyramid approach to subpixel registration based on intensity. IEEE transactions on image processing 7, 27–41 (1998).
16. Dubbs A, Guevara J & Yuste R moco: Fast motion correction for calcium imaging. Frontiers in neuroinformatics 10, 6 (2016).
17. Pnevmatikakis EA & Giovannucci A NoRMCorre: An online algorithm for piecewise rigid motion correction of calcium imaging data. Journal of neuroscience methods 291, 83–94 (2017).
18. Mitani A & Komiyama T Real-time processing of two-photon calcium imaging data including lateral motion artifact correction. Frontiers in neuroinformatics 12, 98 (2018).
19. Mukamel EA, Nimmerjahn A & Schnitzer MJ Automated analysis of cellular signals from large-scale calcium imaging data. Neuron 63, 747–760 (2009).
20. Pachitariu M et al. Suite2p: beyond 10,000 neurons with standard two-photon microscopy. BioRxiv, 061507 (2016).

21. Pnevmatikakis EA et al. Simultaneous denoising, deconvolution, and demixing of calcium imaging data. Neuron 89, 285–299 (2016).
22. Friedrich J, Zhou P & Paninski L Fast online deconvolution of calcium imaging data. PLoS computational biology 13, e1005423 (2017).
23. Keemink SW et al. FISSA: A neuropil decontamination toolbox for calcium imaging signals. Scientific reports 8, 1–12 (2018).
24. Giovannucci A et al. CaImAn an open source tool for scalable calcium imaging data analysis. elife 8, e38173 (2019).
25. Spaen Q et al. HNCcorr: A novel combinatorial approach for cell identification in calcium-imaging movies. eneuro 6, 0304–0318 (2019).
26. Soltanian-Zadeh S, Sahingur K, Blau S, Gong YY & Farsiu S Fast and robust active neuron segmentation in two-photon calcium imaging using spatiotemporal deep learning. P Natl Acad Sci USA 116, 8554–8563, doi:10.1073/pnas.1812995116 (2019).
27. Bao YJ, Soltanian-Zadeh S, Farsiu S & Gong YY Segmentation of neurons from fluorescence calcium recordings beyond real time. Nat Mach Intell 3, 590–600, doi:10.1038/s42256-021-00342-x (2021).
28. Sita L et al. A deep-learning approach for online cell identification and trace extraction in functional two-photon calcium imaging. Nat Commun 13, doi:ARTN 152910.1038/s41467-022-29180-0 (2022).
29. Zhang Y et al. Rapid detection of neurons in widefield calcium imaging datasets after training with synthetic data. Nat Methods 20, 747–754, doi:10.1038/s41592-023-01838-7 (2023).
30. Li X et al. Reinforcing neuron extraction and spike inference in calcium imaging using deep self-supervised denoising. Nature Methods 18, 1395–1400 (2021).
31. Li X et al. Real-time denoising enables high-sensitivity fluorescence time-lapse imaging beyond the shot-noise limit. Nature Biotechnology 41, 282–292 (2022).
32. Lecoq J et al. Removing independent noise in systems neuroscience data using DeepInterpolation. Nature methods 18, 1401–1408 (2021).
33. Zhang J & Ghanem B ISTA-Net: Interpretable optimization-inspired deep network for image compressive sensing. Proceedings of the IEEE conference on computer vision and pattern recognition, 1828–1837 (2018).
34. Hochreiter S & Schmidhuber J Long short-term memory. Neural computation 9, 1735–1780 (1997).
35. Ruder S An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098 (2017).
36. Song A, Gauthier JL, Pillow JW, Tank DW & Charles AS Neural anatomy and optical microscopy (NAOMi) simulation for evaluating calcium imaging methods. Journal of neuroscience methods 358, 109173 (2021).
37. Beck A & Teboulle M A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM journal on imaging sciences 2, 183–202 (2009).
38. Shi X et al. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. Advances in neural information processing systems 28, 802–810 (2015).
39. Rumelhart DE, Hinton GE & Williams RJ Learning representations by back-propagating errors. nature 323, 533–536 (1986).
40. Sundermeyer M, Schlüter R & Ney H LSTM neural networks for language modeling. Thirteenth annual conference of the international speech communication association 65, 194–197 (2012).
41. Ulku I & Akagündüz E A survey on deep learning-based architectures for semantic segmentation on 2d images. Applied Artificial Intelligence 36, 2032924 (2022).
42. Chen L-C, Zhu Y, Papandreou G, Schroff F & Adam H Encoder-decoder with atrous separable convolution for semantic image segmentation. Proceedings of the European conference on computer vision (ECCV), 801–818 (2018).
43. Ronneberger O, Fischer P, & Brox T. U-Net: convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015 Vol. 9351 (eds Navab N, Hornegger J, Wells W & Frangi A) 234–241 (Springer, 2015).

44. Chen Y et al. Brain MRI super resolution using 3D deep densely connected neural networks. 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018), 739–742 (2018).
45. Huang G, Liu Z, Van Der Maaten L & Weinberger KQ Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition, 4700–4708 (2017).
46. Allen Brain Observatory (ABO) calcium imaging dataset. https://github.com/AllenInstitute/AllenSDK/wiki/Use-the-Allen-Brain-Observatory-%E2%80%93-Visual-Coding-on-AWS.
47. Chen T-W et al. Ultrasensitive fluorescent proteins for imaging neuronal activity. Nature 499, 295–300 (2013).
48. Dana H et al. High-performance calcium sensors for imaging activity in neuronal populations and microcompartments. Nature methods 16, 649–657 (2019).
49. Soille P. Morphological Image Analysis: Principles and Applications 105–137 (Springer, 2004). [Google Scholar link]
50. Shattuck DW & Leahy RM BrainSuite: an automated cortical surface identification tool. Medical image analysis 6, 129–142 (2002).
51. Zhang K & Yang W Code for DeepCaImX. Zenodo https://doi.org/10.5281/zenodo.12650420 (2024).

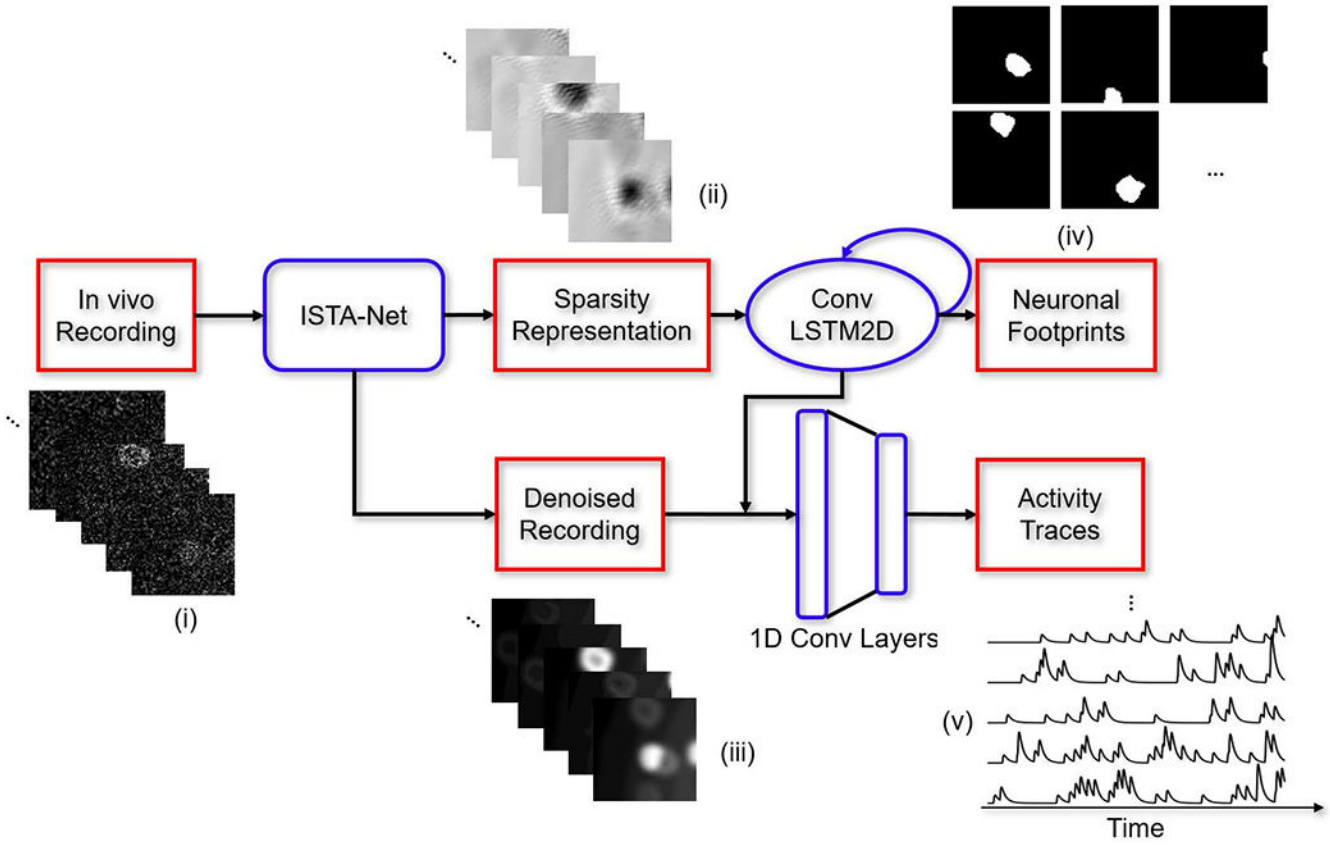


Fig. 1 |. Architecture of DeepCaImX.

DeepCaImX is composed of three modules: ISTA-net, ConvLSTM2D and 1D Convolutional Layers. The raw calcium imaging recording (i) is fed to ISTA-Net, which denoises and removes the background of the recording and transforms it from the spatial domain to a sparse representation. Using the recording in the sparse representation (ii), ConvLSTM2D analyzes the calcium dynamics, and generates an attention probability map for each ROI at each frame as well as the overall ROI segmentation results (iv). The 1D Convolutional Layers then use the attention maps to extract the activity traces of each ROI (v) in the denoised and background-suppressed recording in the spatial domain (iii) output from ISTA-Net.

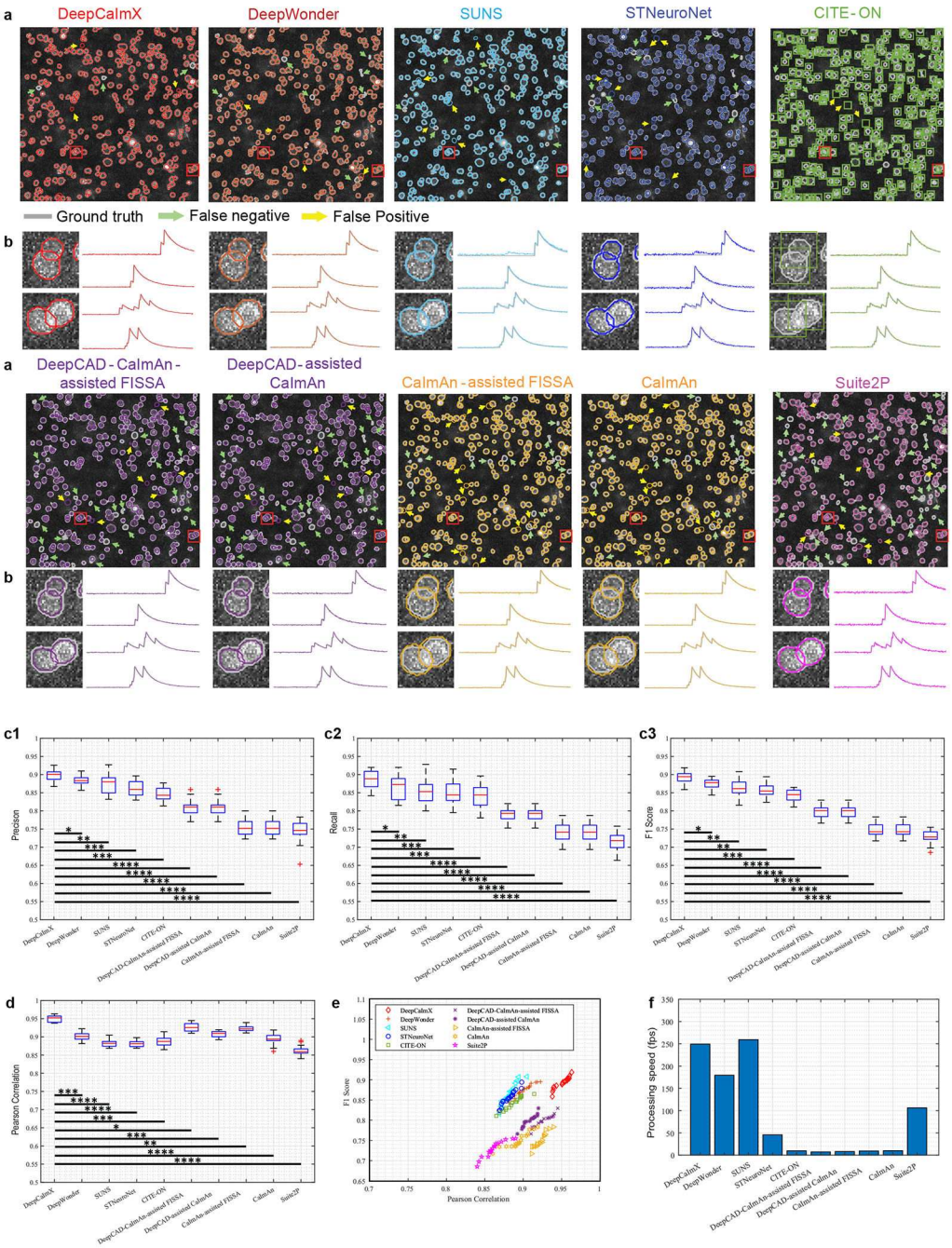


Fig. 2 |. DeepCaImX outperforms other existing methods in neuronal segmentation and trace demixing on simulated data.

a, Segmentations of a simulated sample for DeepCaImX, DeepWonder, SUNS, STNeuroNet, CITE-On, DeepCAD-CaImAn-assisted FISSA, DeepCAD-assisted CaImAn, CaImAn-assisted FISSA, CaImAn, and Suite2p, overlaid on top of the time-series maximum intensity projection of the video. The gray outlines denote the ground truth boundaries of neurons. The color outlines denote the segmentation results of the used method. The yellow and green arrows indicate the false positive and false negative segmented neurons

respectively. **b**, Spatial footprints and temporal activity traces of exemplary neurons from the boxed regions in **a**. **c1-c3**, Recall, precision, and F1 scores of DeepCaImX, DeepWonder, SUNS, STNeuroNet, CITE-On, DeepCAD-CaImAn-assisted FISSA, DeepCAD-assisted CaImAn, CaImAn-assisted FISSA, CaImAn, and Suite2p, for 18 simulated samples, covering all 6 types of calcium indicators (GCaMP6s, 6f and jGCaMP7b, 7c, 7s, 7f). Each sample has a size of 488 (pixels) × 488 (pixels) × 1000 (frames). **d**, Pearson-correlation of extracted activity traces with ground truth traces of common ROIs detected by all methods, for 18 simulated samples. **e**, F1 scores vs Pearson-correlation between extracted activity traces and ground truth traces of common ROIs detected by all the methods, for 18 simulated samples. **f**, Processing speed of different methods. Each frame has 488x488 pixels. Box plot: center bars (red), medians; box edges, first and third quartiles, respectively; whiskers, minimum and maximum; +mark, outliner. *, $p<0.05$; **, $p<0.01$; ***, $p<0.001$; ****, $p<0.0001$, in one-way, two-sided Analysis of Variance (ANOVA), followed by Tukey's Honestly Significant Difference (HSD) test as a post-hoc multiple comparison test.

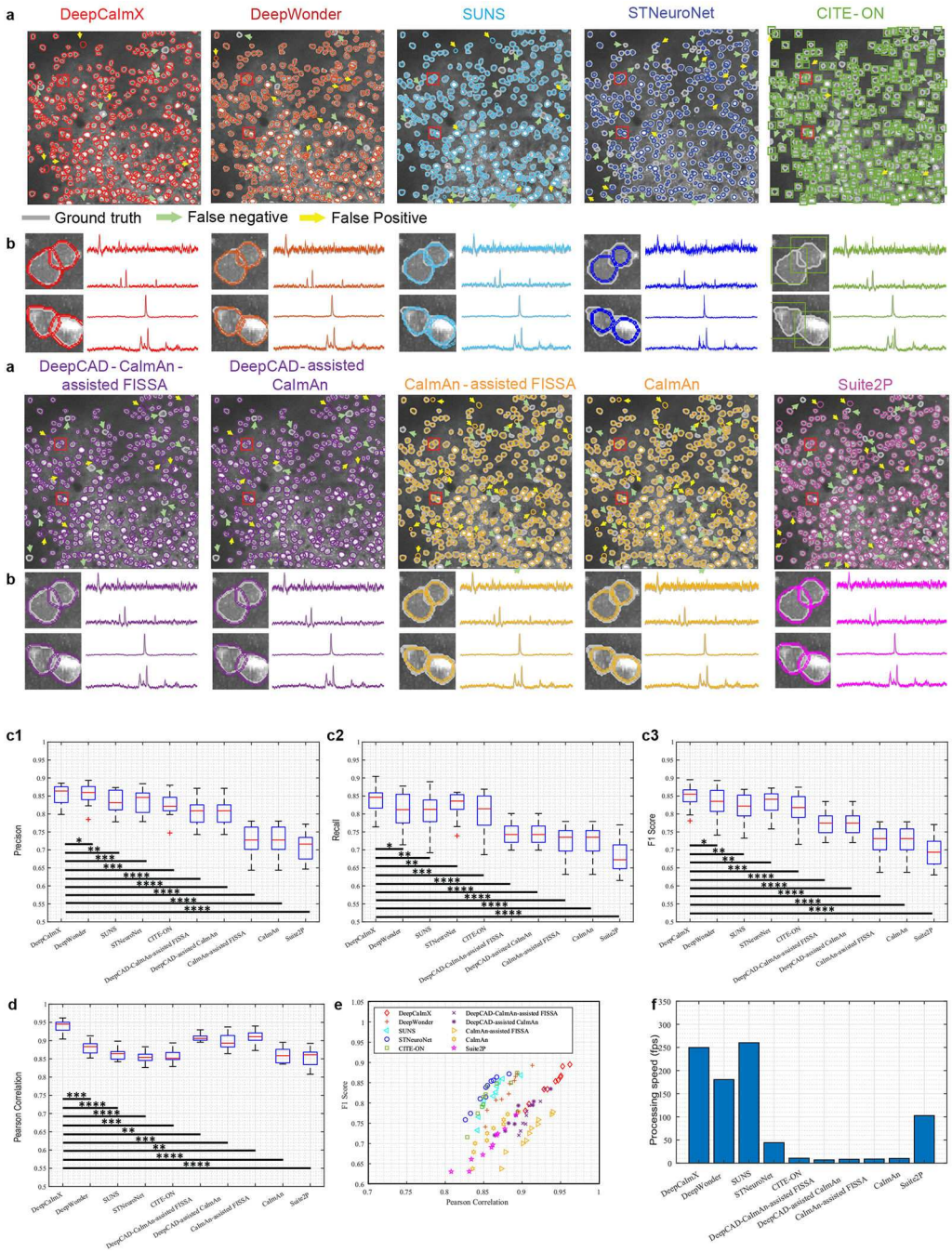


Fig. 3 |. DeepCaImX outperforms existing methods in neuron segmentation and trace extraction on experimental data.

a, Segmentations from an exemplary experimental dataset (ABO 524691284) for DeepCaImX, DeepWonder, SUNS, STNeuroNet, CITE-On, DeepCAD-CaImAn-assisted FISSA, DeepCAD-assisted CaImAn, CaImAn-assisted FISSA, CaImAn, and Suite2p, overlaid on top of the time-series maximum intensity projection of the video. The gray outlines denote the ground truth boundaries of neurons. The color outlines denote the segmentation results of the used method. The yellow and green arrows indicate the

false positive and false negative segmented neurons respectively. **b**, Spatial footprints and temporal activity traces of exemplary neurons from the boxed regions in **a**. **c1-c3**, Recall, precision, and F1 scores of DeepCaImX, DeepWonder, SUNS, STNeuroNet, CITE-On, DeepCAD-CaImAn-assisted FISSA, DeepCAD-assisted CaImAn, CaImAn-assisted FISSA, CaImAn, and Suite2p, for 10 samples over the imaging depth of 175 μm, 275 μm, 375 μm, 550 μm and 625 μm. Each sample has a size of 488 (pixels) × 488 (pixels) × 5000 (frames). **d**, Pearson-correlation of extracted activity traces with ground truth traces of common ROIs detected by all the methods, for 10 samples. **e,** F1 scores vs Pearson-correlation between extracted activity traces and ground truth traces of common ROIs detected by all the methods, for 10 samples. **f**, Processing speed of different methods. Each frame has 488x488 pixels. Box plot: center bars (red), medians; box edges, first and third quartiles, respectively; whiskers, minimum and maximum; +mark, outliner. *, $p<0.05$; **, $p<0.01$; ***, $p<0.001$; ****, $p<0.0001$, in one-way, two-sided Analysis of Variance (ANOVA), followed by Tukey's Honestly Significant Difference (HSD) test as a post-hoc multiple comparison test.

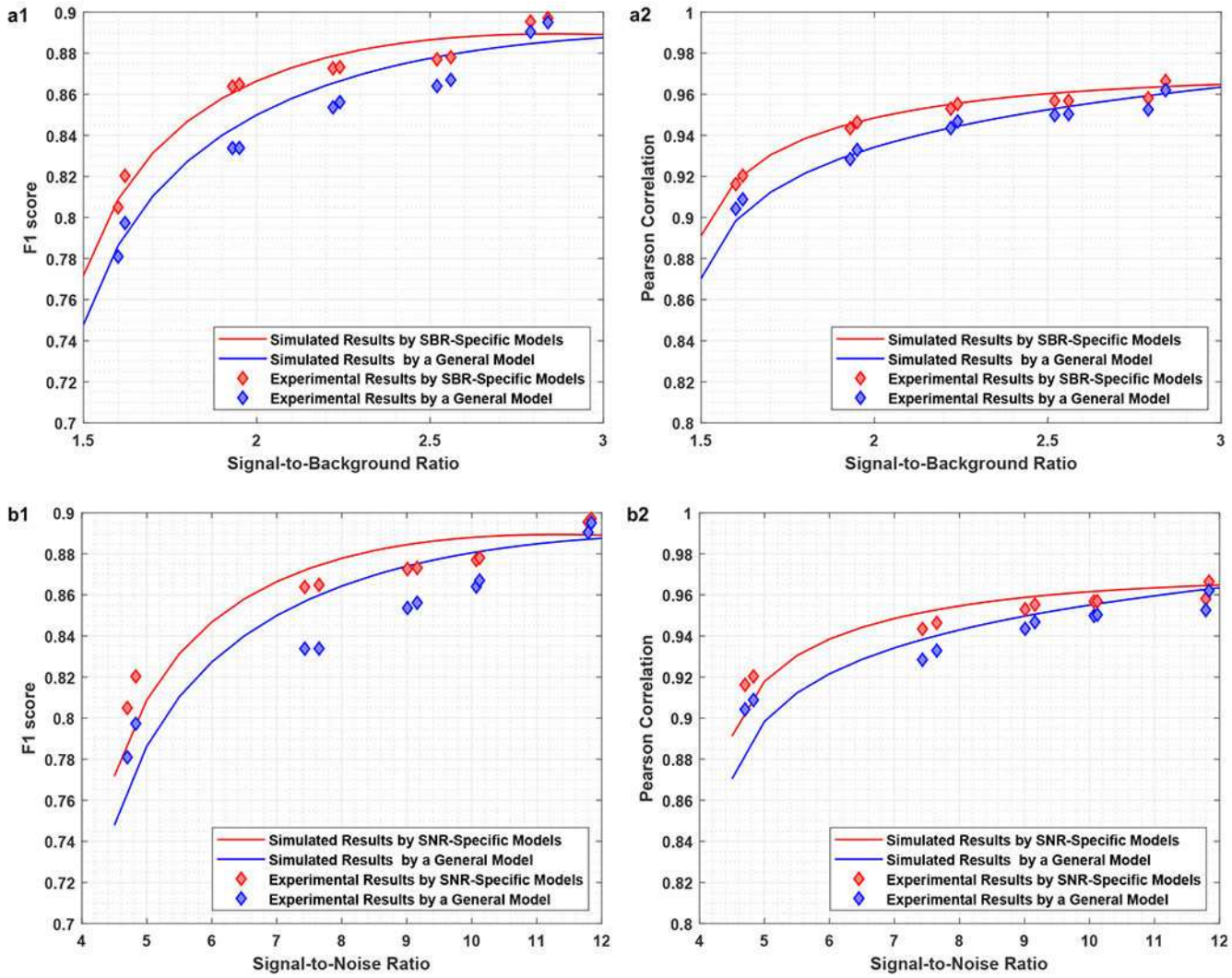


Fig. 4 |. Performance of DeepCaImX on ROI detection and activity trace extraction across different SBRs and SNRs.

a1-a2, F1 score (a1) of the segmentation results in simulated and experimental datasets with different levels of SBR. Pearson correlation (a2) between extracted activity traces and ground truth traces of ROIs detected in simulated and experimental datasets with different levels of SBR. **b1-b2**, F1 score (b1) of the segmentation results in simulated and experimental datasets with different levels of SNR. Pearson correlation (b2) between extracted activity traces and ground truth traces of ROIs detected in simulated and experimental datasets with different levels of SNR. For ABO dataset at image depths of 175 μm, 275 μm, 375 μm, 550 μm and 625 μm, the SBRs are respectively 2.82, 2.54, 2.23, 1.94 and 1.61, and the SNRs are respectively 11.82, 10.09, 9.08, 7.54 and 4.77.