Supplementary Note 1: vTwINS Design and Beam Simulations

In traditional TPM, a large diameter collimated Gaussian beam is centered on the objective back aperture, and a standard (typically diffraction-limited) PSF for two-photon excitation is produced (Fig. 1d left). In contrast, a pair of smaller diameter collimated Gaussian beams with their centers offset from the center of the back aperture produce a pair of elongated arms in the PSF that cross at the focal plane, producing an X-shape (Fig. 1d center). The smaller the diameter of each beam, which reduces the effective NA, the more each arm becomes elongated. Increasing the separation between the two beams at the objective back aperture increases the angle of intersection of the two arms. If the incident beams are slightly divergent (or convergent), the position of the crossing point of the two beams shifts along the optical axis relative to the focal point of the objective (Fig. 1d right), eventually producing a V-shape for vTwINS, with the wider opening either pointing up (divergent beams) or down (convergent beams). To produce a Bessel beam vTwINS PSF, rings of illumination are used at the objective back aperture [1] instead of Gaussian beams, but, otherwise, the same principles apply.

The beam-shaping module (Fig. 1c) was designed with three parallel beam paths, one each back aperture illumination profile described above. The vTwINS paths are separated into three parts: 1. beam-shaping for Gaussian or Bessel beams, 2. adjusting the V offset with a telescope 3. Splitting the beams in two with polarizing blocks and setting the V angle

1. An variable zoom telescope was used to shape the Gaussian beam path, while the Bessel beam vTwINS used an axicon-lens combination that formed the ring-shaped spatial profile.

2. Beam divergence, used to control where the two beams cross (and forming either a V or inverted V), was controlled by an adjustable telescope placed before the beam splitter

3. A calcite block was used to split a single incident beam into the two spatially separated beams necessary for vTwINS. Input polarization, controlled by a half-wave plate, was used to equalize power in each of the two beams. The angle between the two arms of the PSF is scaled by the zoom factor in a subsequent fixed zoom telescope.

Both Bessel beams and Gaussian beams were used with vTwINS, each with their own advantages. To explore the relative merits of Gaussian beam versus Bessel beam PSFs, PSF simulations (Supplementary Fig. 1) were generated in ZEMAX using the previously described optical setup. The Bessel beam PSFs were generated with a 179° BK7 Axicon and 150/200mm achromat lens pair. The Gaussian beam PSF corresponds to a diffraction-limited 0.175NA Gaussian beam. Gaussian beams are more easily implemented and have higher overall two-photon excitation for a given incident power. The smaller profile of the Gaussian beams on the back aperture allows for a larger allowable V angle range for the vTwINS PSF as compared to Bessel beams. Bessel beams have a controllable axial profile [1], minimizing excitation outside the vTwINS imaging volume and can be used to compensate for additional scattering in the deeper regions of the vTwINS imaging volume. While higher lateral resolutions are achieved with Bessel beams, improving the lateral resolution comes at the cost of total two-photon excitation efficiency. Additionally, improved lateral resolution is not advantageous in vTwINS, as the uniqueness of spatial profiles makes sharp cell features unnecessary for resolving individual cells.

Supplementary Note 2: Fluorescent Bead Sample and Measured PSFs

As an initial test of vTwINS we imaged 1 μm green fluorescent beads (L1030, Sigma) embedded in a 1% agarose gel. The beads were embedded at random locations, creating an off-grid set of positions. The exact bead positions were determined via a diffraction-limited two-photon multi-plane volumetric scan (z-stack). vTwINS was then used to image the same volume with a single scan (one image). Each bead appears in the vTwINS projection image as a pair of dots (Supplementary Fig. 2); lines drawn between all pairs are parallel and aligned with the direction of the vTwINS PSF in the sample. The distance between dot pairs varies with the bead's depth in the volume.

Using the single vTwINS image, SCISM was used to automatically locate the spatial profiles for each bead. The found profiles were used in turn to infer each bead's 3D coordinates in the sample, using the vTwINS relationship between depth and inter-image distance. Beads at the edge of the imaged volume with only one projection into the vTwINS image were discarded as the depth location could not be ascertained. Using only the vTwINS image and the known shape of the PSF to infer each bead's 3D coordinates in the sample produces average errors of 1.4$\pm$1.3 μm (N=31) in depth, 1.5$\pm$1.3 μm (N=31) in the fast-scan direction, 1.2$\pm$1.0 μm (N=31)in the slow-scan direction and an average total localization error of 2.7$\pm$1.6 μm (N=31). The accuracy of recovered positions is well within the $\approx$10 μm (N=31) average size of a neuronal cell body in the mammalian brain, demonstrating that vTwINS, in practice, preserves the necessary information to disambiguate cell bodies at different depths.

Z-stacks of these fluorescent beads (1 μm step size) were taken to measure the vTwINS PSFs for each set of experiments. The axial length of PSFs were measured by averaging the fluorescence intensity at each slice of the z-stack. The averaged fluorescence signal was used to calculate the FWHM and 1/e full-width axial lengths. For *in vivo* experiments, we found that the 1/e full-width was a reasonable cutoff for acceptable imaging quality. The PSFs were not measured *in vivo*, although this can be done to correct for index mismatch and scattering if higher accuracy is desired.

Supplementary Note 3: Spatial Overlap Error Estimation

Spatial profiles extracted using SCISM from the datasets in Supplementary Fig.7a,b were used to estimate the maximum level of spatial overlap present in vTwINS and with an single axially extended beam. Spatial overlap was calculated as the fraction of overlapping pixels between two spatial profiles (pixels were included if they had a profile weight greater than 10% of the maximum weight in the spatial profile). The maximum spatial overlap was calculated as the max of the set of spatial overlaps each profile had with all other profiles. The distribution of maximum spatial overlaps for vTwINS are shown in Supplementary Fig. 3a,b. The spatial overlap in the case of a single axially extended beam was calculated by considering only half of the spatial profiles isolated using SCISM. We isolated the left half of the spatial profiles and then calculated the distribution of maximum spatial overlap for all cells, plotted in Supplementary Fig. 3c,d. The spatial overlap of vTwINS profiles is substantially lower than the spatial overlap of a single axially extended beam. It is unclear, however, what level of spatial overlap prevents successful demixing.

The performance of CNMF with a single axially extended beam was compared to SCISM with a vTwINS PSF using the sequential alternating beams dataset (Supplementary Fig. 13). CNMF was run on the frames of only one beam, corresponding to a dataset with a single axially extended beam, and was compared the output to the SCISM output of an interleaved version of the same dataset, approximating a vTwINS movie (see Methods). Spatial profiles from CNMF used on a single beam were paired off with vTwINS spatial profiles based upon their time traces' Pearson's correlation. Among the matched profiles, there were cases where vTwINS profiles were not paired with their most highly correlated CNMF trace (Supplementary Fig.4a). In this example, the high spatial overlap of the cells caused CNMF to merge the activity of two cells into a single profile, preventing accurate recovery of neural data.

The performance of CNMF with vTwINS data was evaluated using the sequential alternating beams dataset (Supplementary Fig. 13). CNMF was run on each of the two beams separately, and also in a vTwINS-equivalent interleaved movie (see Methods). Spatial profiles generated by the interleaved vTwINS movie were paired off with the set of spatial profiles generated by each of the two single axially extended beam movies based on their time traces' Pearson's correlation. Among the matched profiles, there were cases where single axially extended beam profiles were not paired with their most highly correlated interleaved vTwINS profile (Supplementary Fig. 4b). This example suggests that the spatial overlap between the axial projections of two different cells prevented CNMF, which does not have a prior on paired spatial profile shapes, from correctly attributing the neural activity to two cells.

Supplementary Note 4: Noise and SNR calculations

We calculate the signal-to-noise ratio (SNR) and peak signal-to-noise ratio (PSNR) for each spatial profile using the least-square time-trace estimates $\widehat{\boldsymbol{S}}$ calculated from

$$\left\{\widehat{\boldsymbol{S}}, \widehat{\boldsymbol{S}}_{bg}\right\} = \arg\min_{\boldsymbol{S},\boldsymbol{S}_{bg}} \left[\left\|\boldsymbol{Y} - \widehat{\boldsymbol{X}}\boldsymbol{S}^T - \widehat{\boldsymbol{X}}_{bg}\boldsymbol{S}_{bg}^T\right\|_F^2\right].$$

Assuming that each column $\boldsymbol{s}$ of $\widehat{\boldsymbol{S}}$ contains both a signal and a noise component, i.e.

$$\boldsymbol{s} = \boldsymbol{s}_{signal} + \boldsymbol{s}_{noise}.$$

then we use the definitions of SNR and PSNR given by

$$\mathrm{SNR} = \frac{\mathrm{Var}(\boldsymbol{s}_{signal})}{\mathrm{Var}(\boldsymbol{s}_{noise})} = \frac{\mathrm{Var}(\boldsymbol{s}) - \sigma_{noise}^2}{\sigma_{noise}^2},$$

where $\sigma_{noise}^2 = \mathrm{Var}(\boldsymbol{s}_{noise})$, and

$$\mathrm{PSNR} = \frac{\|\boldsymbol{s}\|_\infty^2}{\sigma_{noise}^2}.$$

The noise variance for the SNR and PSNR calculations, σ_{noise}^2, is calculated from the σ estimate used to determine significant transients. For each least-squares calcium trace, $\boldsymbol{s}$, the FHWM of the calcium trace histogram is used to estimate $\sigma \approx FWHM/2.35$, approximating the distribution as a Gaussian. Transients are described as significant if they exceed a 3σ cutoff above the center of the distribution.

Supplementary Note 5: Phototoxicity

Phototoxicity is an important concern of live optical imaging techniques, especially for large-scale or volumetric imaging techniques where high phototoxicity might be expected due to high total power. In TPM, phototoxicity is nonlinear and has been shown to scale as $\int I(\vec{x},t)^{2.5}dVdt$, where $I(\vec{x},t)$ is the intensity distribution, V is the excited volume, and t is time [2, 3]. The intensity exponent reflects that nonlinear photodamage has both second and third order contributions. Here, we use this relationship to compare the accumulated phototoxicity in vTwINS imaging with high-NA TPM.

In high-NA two photon imaging, studies have routinely imaged L2/3 in mouse cortex (300×150 μm FOV) with 80 mW [4] and CA1 in mouse hippocampus (200×100 μm FOV) with 50 mW [5] in daily imaging sessions without noticeable phototoxicity. In vTwINS, we have not observed photobleaching or phototoxicity in mice with hour-long imaging sessions at up to 200 mW average power (100 mW per beam).

To calculate the relative phototoxicity between the two conditions, we first describe the PSF within the sample. For a highly scattering tissue like mouse cortex, scattering of the focused beam reduces the effective power at the sample with a decay length of ≈ 140 μm and distortions in the wavefront degrade the PSF. For a 0.8 NA PSF, this is reduced to an effective 0.5 NA [6].

For equivalent laser parameters, the total integrated phototoxicity through a given cell is given by the total integral of the raster-scanned PSF through the cell volume:

$$\begin{aligned}
&\int\left[\int I(\vec{x},t)^{2.5}dt\right]dV \\
&=\int\left[\int I_g(x,y,z)^{2.5}*R(x,y,t)dt\right]dV \\
&=\int\left[I_g(x,y,z)^{2.5}*\int R(x,y,t)dt\right]dV
\end{aligned}$$

Where $I_g(x,y,z)$ is the intensity distribution of a Gaussian beam, which is raised to the $2.5th$ power for phototoxicity and convolved with $R(x,y,t)$, the raster scanning function. As I_g has no time dependence, we may integrate $R(x,y,t)$ over time, assuming laser pulses are uniformly distributed in time over space:

$$\int R(x,y,t)dt\approx\begin{cases}\frac{1}{\Delta x\Delta y} & |x|\leqslant\Delta x/2 \quad and \quad |y|\leqslant\Delta y/2\\ 0 & otherwise\end{cases}$$

where Δx and Δy are the raster scanning dimensions and the factor of $(1)/(\Delta x\Delta y)$ reflects the duty cycle over the scanned area. Using the equation for the intensity of a Gaussian beam:

$$=n\int_{cell}\left(\frac{P_0}{w(0)^2}(\frac{w(0)}{w(z)})^2e^{-2r^2/w(z)^2}\right)^{2.5}*\int R(x,y,t)dtdV$$

where a factor n was added for the number of imaging beams and where $w(z)=\sqrt{(\lambda/\pi\mathrm{NA})^2+(z\mathrm{NA})^2}$. Using the expression for $\int R(x,y,t)dt$, we obtain:

$$=n\int_{cell}\left(\frac{P_0}{w(z)^2}e^{-2r^2/w(z)^2}\right)^{2.5}*\left\{\begin{array}{cc}\frac{1}{\Delta x\Delta y} & |x|\leqslant\Delta x/2 \quad and \quad |y|\leqslant\Delta y/2\\ 0 & otherwise\end{array}\right\}dV$$

As the size of the PSF is much smaller than that of the scanned region, we can approximate the convolution as an integral over the xy plane

$$
\begin{aligned}
&= \frac{n}{\Delta x \Delta y} \int_{cell} 2\pi \int_0^\infty \left(\frac{P_0}{w(z)^2} e^{-2r^2/w(z)^2}\right)^{2.5} r dr dV \\
&= \frac{n}{\Delta x \Delta y} \int_{cell} \frac{\pi P_0^{2.5}}{5 w(z)^3} dV \\
&= \frac{\pi n P_0^{2.5}}{5 \Delta x \Delta y} \int_{-r_0}^{r_0} \frac{\pi (r_0^2 - z^2)}{w(z)^3} dz
\end{aligned}
$$

where $r_0 = 6$ μm is the radius of the typical cell. Here, the cell was placed at $z = 0$, which maximizes the total integrated phototoxicity.

$$
\begin{aligned}
&= \frac{\pi^2 n P_0^{2.5}}{5 \Delta x \Delta y} \int_{-r_0}^{r_0} \frac{(r_0^2 - z^2)}{((\lambda/\pi NA)^2 + (zNA)^2)^{1.5}} dz \\
&= \frac{\pi^2 n P_0^{2.5}}{5 \Delta x \Delta y} \left(\frac{2 r_0 \pi^2 w(r_0)}{\lambda^2} + \frac{1}{NA^3} log\left(\frac{w(r_0) - NA r_0}{w(r_0) + NA r_0}\right) \right)
\end{aligned}
$$

Evaluating this for high NA TPM in the conditions above for cortical imaging ($\Delta x = 300$ μm, $\Delta y = 150$ μm, $P_0 = 80$ mW, NA = 0.5, n = 1) obtains 980 AU, and for CA1 imaging ($\Delta x = 200$ μm, $\Delta y = 100$ μm, $P_0 = 50$ mW, NA = 0.5, n = 1) we obtain 680 AU. When evaluating this for vTwINS ($\Delta x = 500$ μm, $\Delta y = 500$ μm, $P_0 = 100$ mW, NA = 0.175, n = 2), we obtain 87 AU. Given typical imaging parameters, we find that the total accumulated phototoxicity in cells being imaged in vTwINS is lower by an order of magnitude than what is typically used during high NA TPM.

An alternative calculation of the nonlinear phototoxicity relies upon the peak volumetric excitation [7]. The power density of a 50 μm vTwINS PSF (100 mW for each half, approximated as a 50 μm axial, 1.71 μm lateral ellipsoid) is 100 mW/(76μm^3) = 1.3 mW/μm^3 (for each half). For a 0.8 NA (effectively 0.5 NA, 6 μm axial, 0.66 μm lateral ellipsoid) Gaussian PSF, this power density would be comparable to a (1.3 mW/μm^3)(1.37μm^3) $\approx$2 mW laser power. Consequently, the peak excitation at any given location while imaging during vTwINS is much lower relative to high NA imaging configurations. This is comparable to the effect described in [7], where a lower excitation NA allows for the use of much higher excitation power.

Supplementary Note 6: vTwINS Orthogonal Matching Pursuit

In this section, we describe the mathematical details of the vTwINS Sparse Convolutional Iterative Shape Matching (SCISM) demixing algorithm. Let $\boldsymbol{Y} \in \mathbb{R}^{N \times T}$ denote the calcium video sequence, $\boldsymbol{X} \in \mathbb{R}^{N \times K}$ denote the neural spatial components (spatial profiles), and $\boldsymbol{S} \in \mathbb{R}^{T \times K}$ denote the neural temporal activity traces, where N is the number of pixels in each image, T is the number of images (or time points), and K is the number of neurons. Thus, the columns of $\boldsymbol{Y}$ represent single frames of the video, the columns of $\boldsymbol{X}$ represent individual spatial profiles, and the columns of $\boldsymbol{S}$ represent temporal activity traces of single neurons. We model background activity with a set of B background components $\boldsymbol{X}_{bg} \in \mathbb{R}^{N \times B}$ and denote the (inferred) background temporal activity $\boldsymbol{S}_{bg} \in \mathbb{R}^{T \times B}$.

Our algorithm is designed to exploit *a priori* knowledge of both the spatial profile shapes as well as neural firing statistics. Specifically, the algorithm seeks to factor the full movie matrix $\boldsymbol{Y}$ into the set of spatial profiles $\boldsymbol{X}$ and time-traces $\boldsymbol{S}$ such that

1. The sum of outer products of spatial profiles and time traces explains the observed data ($\boldsymbol{Y} \approx \boldsymbol{X}\boldsymbol{S}^T$).
2. The time-traces $\boldsymbol{S}$ are sparse in time.
3. The spatial profiles are shaped like pairs of neuronal somata (disks or annuli), offset horizontally by a small separation distance. The dark center in each soma is due to the lack of GCaMP6f in the nucleus.
4. Few latent sources (active neurons) relative to the size of the data generate activity in the observed data, making the fluorescence movie low-rank. This constraint captures the physical density constraints on neuron tissue.

The optimization program that includes all these terms is

$$\left\{\widehat{\boldsymbol{X}}, \widehat{\boldsymbol{S}}, \widehat{\boldsymbol{X}}_{bg}, \widehat{\boldsymbol{S}}_{bg}\right\} = \arg \min_{\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{X}_{bg}, \boldsymbol{S}_{bg} \geqslant 0} \left[\left\| \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{S}^T - \boldsymbol{X}_{bg}\boldsymbol{S}_{bg}^T \right\|_F^2 + \lambda_d \|\boldsymbol{X} - \boldsymbol{D}\|_F^2 + \sum_k \left(\lambda_{gs} \|\boldsymbol{s}_k\|_2 + \lambda_{sp} \|\boldsymbol{s}_k\|_1\right) \right] \quad (1)$$

where $\boldsymbol{s}_k$ is the k^{th} column of $\boldsymbol{S}$, representing the activity of neuron k, $\|\boldsymbol{Z}\|_F^2 = \sum_{i,j} Z_{i,j}^2$ is the squared-Frobenius norm, $\boldsymbol{D}$ is a matrix whose columns represent all possible expected neural spatial profile shapes, λ_d is the trade-off parameter for penalizing the deviation of spatial profile shapes $\boldsymbol{X}$ from the idealized shapes in $\boldsymbol{D}$, λ_{gs} is the group-sparse penalization parameter for ensuring that not all spatial profiles are active and λ_{sp} is the penalization parameter than ensures the time traces are sparse. Each column $\boldsymbol{d}_k$ of $\boldsymbol{D}$ represents the expected spatial profile for a neuron at one volumetric neural location. We set the spatial profiles $\boldsymbol{d}_k$ as annuli separated by a depth-dependent distance (Supplementary Fig. 5a), where the annuli were modeled as the difference of two Gaussian functions, separated by a distance

$$d_k(i,j) = e^{-\frac{(i-i_x-\Delta/2)^2+(j-j_y)^2}{\sigma_{out}^2}} - Ae^{-\frac{(i-i_x-\Delta/2)^2+(j-j_y)^2}{\sigma_{in}^2}} + e^{-\frac{(i-i_x+\Delta/2)^2+(j-j_y)^2}{\sigma_{out}^2}} - Ae^{-\frac{(i-i_x+\Delta/2)^2+(j-j_y)^2}{\sigma_{in}^2}}$$

For all datasets analyzed here, the annuli were set to have $\sigma_{out} = 2$ pixels and $\sigma_{in} = 0.84$ pixels, the center amplitude depression was set to $A = 0.7$, and (i_x, j_y) simply indicate the pixel which $\boldsymbol{d}_k$ is centered around. We used 10 different inter-image distances, Δ, equally spaced between 21.4 μm to 92.4 μm for full FOV V1 data spanned, 18.4 μm and 51.4 μm for half FOV V1 data, and 14.6 μm to 56.8 μm for full FOV CA1 data. In total, the number of columns of $\boldsymbol{D}$ is the number of pixels N (all potential spatial locations) times the number of inter-image distances K ($\boldsymbol{D} \in \mathbb{R}^{N \times NK}$). This matrix, however, never needs to be constructed, as any the spatial invariance of the neural profiles permits the use of convolution operations. The parameters used for our analysis reflects the particulars of our microscope setup (i.e. zoom, beam angle setting etc.) and should be modified to fit the expected statistics of any new dataset.

The optimization program in Equation (1) results from modeling the measurement noise as Gaussian, and placing appropriate sparsity- and shape-penalizing priors on the spatial profiles and transients. The measurement model and Gaussian prior over the spatial profiles $\boldsymbol{X}$ are given by

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{S}^T + \boldsymbol{X}_{bg}\boldsymbol{S}_{bg}^T + \boldsymbol{E}, \qquad E_{i,j} \sim \mathcal{N}(0, \sigma^2)$$
$$\boldsymbol{x}_k \sim \mathcal{N}(\boldsymbol{d}_k, \sigma_p^2 \boldsymbol{I}),$$

where the non-zero mean of the Gaussian prior over spatial profiles induces the expected spatial structure. The prior over time traces $\boldsymbol{s}_k$

$$p(\boldsymbol{s}_k) \propto e^{-\gamma_1 \|\boldsymbol{s}_k\|_2 - \gamma_2 \|\boldsymbol{s}_k\|_1},$$

includes two terms penalizing both overall sparsity and group sparsity (each neural trace being a group). In terms of the model parameters, the trade-off parameters in Equation 1 are $\lambda_d = \sigma^2/\sigma_d^2$, $\lambda_{gs} = \gamma_1\sigma^2$, and $\lambda_{sp} = \gamma_2\sigma^2$. No specific prior was placed on either the background shape or its temporal fluctuations.

Direct optimization of Equation (1) can be inefficient due to the problem size and the large search space (potential spatial profiles). We thus approximated a solution to Equation (1) with a greedy, iterative approach wherein spatial profiles are sequentially determined. Our method iterates between finding the best element of $\boldsymbol{D}$ that approximates $\boldsymbol{Y}$ given the sparsity constraints and updating that profile to the data. The first step sets $\boldsymbol{X} = \boldsymbol{D}$ and solves for the best single trace to approximate $\boldsymbol{Y}$ (solving the first and third terms). The shape refinement step then uses the first two terms with the newly found time-trace to allow the spatial profile $\boldsymbol{x}_k$ to deviate from its mean $\boldsymbol{d}_k$. SCISM is in essence a modification of the orthogonal matching pursuit (OMP) method for greedy sparse signal estimation [8, 9]. Our method extends OMP by including an additional temporal sparsity penalty and a shape refinement step that allows for deviations from the stereotyped neuronal shapes (traditional OMP assumes a fixed dictionary of features).

We initialized our algorithm by estimating the background spatial profile, $\boldsymbol{X}_{bg}$ using the normalized temporal median of the pre-processed motion-corrected video sequence and $\boldsymbol{S}_{bg}$ as its least-squares time course

$$\widehat{\boldsymbol{X}}_{bg} = \frac{\text{Median}(\boldsymbol{y}_t)}{\|\text{Median}(\boldsymbol{y}_t)\|_2}, \qquad \widehat{\boldsymbol{S}}_{bg} = \widehat{\boldsymbol{X}}_{bg}^{+}\boldsymbol{Y},$$

where $\boldsymbol{X}^+$ denotes the pseudo-inverse of $\boldsymbol{X}$. In the case of shorter video sequences (20000 frames or less) we only used a single background spatial profile ($B = 1$). For longer video sequences a background spatial profile was added for each 5000 consecutive and non-overlapping frames, which

allowed the background to change over the course of the video sequence (e.g. due to slow axial drift). The *residual* movie $\boldsymbol{R}$ was then initialized at the first step to the median-subtracted full movie $\boldsymbol{Y}$

$$\boldsymbol{R} = \boldsymbol{Y} - \widehat{\boldsymbol{X}}_{bg}\widehat{\boldsymbol{S}}_{bg}^{T}.$$

The algorithm (summarized graphically in Fig. 3, Supplementary Fig. 5 and algorithmically in Alg. 1) begins each iteration by seeking the stereotyped annuli pair that had the largest correlation with the residual movie $\boldsymbol{R}$. Specifically, the algorithm seeks the index k and the corresponding pair $\boldsymbol{d}_k$ with the largest value v_k calculated as,

$$v_k = \sum_t T_\lambda \left(\boldsymbol{d}_k^T \boldsymbol{r}_t\right)^2, \tag{2}$$

where $\boldsymbol{T}_\lambda$ is a soft thresholding function restricted to positive values

$$T_\lambda(x) = \begin{cases} x - \lambda & x \geqslant \lambda \\ 0 & x < \lambda \end{cases}, \tag{3}$$

and $\boldsymbol{r}_t$ are the columns of $\boldsymbol{R}$ (the frames of the residual video). v_k estimates the total energy of the estimated time trace $\boldsymbol{s}_k$ that minimized Equation (1) conditioned on $\boldsymbol{x}_k = \boldsymbol{d}_k$, and all past profiles and time traces being fixed. The thresholding operation induces temporal sparsity (the last term in Eqn. (1)), and prevents noise accumulation over long videos from dominating the values of v_k. Thus even very sparsely firing neurons can be identified, provided they fluoresce above the noise floor. Because the noise floor is not spatially constant, we set the sparsity penalization parameter λ to be a function of the local statistics affecting each potential spatial profile shape. Specifically, we set λ to be proportional to the 99^{th} percentile of the residual projected into the stereotyped shapes,

$$\lambda_k = 0.05 * p_{0.99}(\boldsymbol{d}_k^T \boldsymbol{r}_t). \tag{4}$$

This local parameter setting measures the potential brightness at each location. As brighter locations have higher backgrounds and higher noise levels, λ is thus set higher at these locations.

After calculating v_k, the stereotyped spatial profile $\boldsymbol{d}_{\widehat{k}}$ at $\widehat{k} = \arg\max_k(v_k)$ is added to the set of spatial profiles. As $\boldsymbol{d}_k$ only approximates the profile shape, a spatial profile that balances the observed data and prior shape information is obtained using a shape refinement step. The shape refinement step estimates $\boldsymbol{x}_k$ from $\boldsymbol{R}$ and $\boldsymbol{d}_k$ as

$$\widehat{\boldsymbol{x}}_k = \frac{1}{N}\sum_t \mathcal{M}\left(\boldsymbol{r}_t\right) \frac{T_\lambda(\boldsymbol{r}_t^T \boldsymbol{d}_k)}{\|\mathcal{M}\left(\boldsymbol{r}_t\right)\|_2} \tag{5}$$

where $\boldsymbol{M}(\cdot)$ is a mask that restricts the averaged frames to the location of $\boldsymbol{d}_k$ (thereby preventing spurious activity from across the video from being included in the spatial profile $\boldsymbol{x}_k$). The normalization by the magnitude of $\boldsymbol{r}_t$ prevented spurious high activity frames, where the activity may not come from that particular neuron, from dominating the average and corrupting the results. In terms of the original cost function, this essentially prevents contributions from yet-to-be located neurons from influencing the spatial profile of the current neuron. While SCISM could be modified to refine all past spatial profiles at each iteration in order incorporate the new profile, such an extension is not explored here.

Given the updated spatial profile list, the time traces $\boldsymbol{S}$ and $\boldsymbol{s}_{bg}$ are obtained via non-negative LASSO

$$\left\{\widehat{\boldsymbol{S}}, \widehat{\boldsymbol{S}}_{bg}\right\} = \arg\min_{\boldsymbol{S}, \boldsymbol{S}_{bg} \geqslant 0} \left[\left\| \boldsymbol{Y} - \widehat{\boldsymbol{X}} \boldsymbol{S}^T - \widehat{\boldsymbol{X}}_{bg} \boldsymbol{S}_{bg}^T \right\|_F^2 + \lambda_{sp} \sum_k \|\boldsymbol{s}_k\|_1 \right] \tag{6}$$

and the residual movie is updated as

$$\boldsymbol{R} = \boldsymbol{Y} - \widehat{\boldsymbol{X}} \widehat{\boldsymbol{S}}^T - \widehat{\boldsymbol{X}}_{bg} \widehat{\boldsymbol{S}}_{bg}^T.$$

The algorithm then repeats, using the new residual to find the next neural spatial profile, starting again from Equation (2).

We ran SCISM until either a pre-set number of spatial profiles was found, or the activity trace for the most recently found spatial profile was essentially zero. Ideally, however, SCISM would iterate until the recovered spatial profiles no longer resemble neurons. While our neural activity-based criterion attempted to determine if newly found spatial profiles represented neurons, more sophisticated methods would increase accuracy. Since testing if a spatial profile represents a neuron is still an open problem [10], one potential approach is to manually check new spatial profiles as they are found and manually stop when newer profiles are deemed to no longer be capturing neural activity. An example of SCISM processing vTwINS data is provided in Supplementary Video 7.

Once the algorithm is ended, the full-temporal resolution time-traces is obtained via non-negative LASSO (Eqn. (6)) with the non-temporally averaged data in place of $\boldsymbol{Y}$.

Algorithm 1 SCISM algorithm for locating pairs of neuronal images in volumetric calcium data.

Set λ_1, λ_2 and K or s_0
Set $m = 1$
Initialize $\widehat{\boldsymbol{X}}_{bg} = \frac{\text{Median}(\boldsymbol{y}_t)}{\|\text{Median}(\boldsymbol{y}_t)\|_2}$, $\widehat{\boldsymbol{S}}_{bg} = \widehat{\boldsymbol{X}}_{bg}^{+} \boldsymbol{Y}$
$\boldsymbol{R} = \boldsymbol{Y} - \widehat{\boldsymbol{X}}_{bg} \widehat{\boldsymbol{X}}_{bg}^{+} \boldsymbol{Y}$
repeat
  $v_l = \sum_t T_{\lambda_1} \left(\boldsymbol{d}_l^T \boldsymbol{r}_t \right)^2$
  $k = \arg\max_l v_l$
  $\widehat{\boldsymbol{x}}_k = \frac{1}{N} \sum_t \mathcal{M}\left(\boldsymbol{r}_t\right) \frac{T_{\lambda_2}(\boldsymbol{r}_t^T \boldsymbol{d}_k)}{\|\mathcal{M}(\boldsymbol{r}_t)\|_2}$
  $\left\{\widehat{\boldsymbol{S}}, \boldsymbol{S}_{bg}\right\} = \arg\min_{\boldsymbol{S}, \boldsymbol{S}_{bg} \geqslant 0} \left[\|\boldsymbol{Y} - \widehat{\boldsymbol{X}} \boldsymbol{S}^T - \widehat{\boldsymbol{X}}_{bg} \boldsymbol{S}_{bg}^T\|_F^2 + \lambda_{sp} \sum_k \|\boldsymbol{s}_k\|_1 \right]$
  $\boldsymbol{R} = \boldsymbol{Y} - \widehat{\boldsymbol{X}} \widehat{\boldsymbol{S}}^T - \widehat{\boldsymbol{X}}_{bg} \widehat{\boldsymbol{S}}_{bg}^T$
  $m = m + 1$
until $\min_k \|\boldsymbol{s}_k\|_2^2 \leqslant s_0$ OR $m >= K$
Output $\widehat{\boldsymbol{X}}, \widehat{\boldsymbol{S}}, \widehat{\boldsymbol{X}}_{bg}, \widehat{\boldsymbol{S}}_{bg}$

To improve the computational efficiency of our method, we introduced two optional modifications to the algorithm's order of computations. First, because inner products of distant spatial profile shapes are nearly independent, multiple new spatial profiles can be selected at each iteration by seeking multiple, well-separated, local maxima of v_t. Second, calculating all inner products with the residual at each iteration can be computationally expensive (essentially K 3D convolutions between each $\boldsymbol{d}_k$ and the data). For small-to-medium sized datasets, we offset some of the computational burden, at the cost of additional memory, by using the linearity of the inner product. Using the

reformulation

$$(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{S}^T)^T\boldsymbol{d}_k = \boldsymbol{Y}^T\boldsymbol{d}_k - \boldsymbol{S}\boldsymbol{X}^T\boldsymbol{d}_k,$$

the algorithm could pre-calculate the inner products with the data ($\boldsymbol{Y} * \boldsymbol{d}_k$) and the spatial profiles ($\boldsymbol{X}^T\boldsymbol{d}_k$), and the inner products with the residual were then calculated via a small number of outer products $\boldsymbol{S}(\boldsymbol{X}^T\boldsymbol{d}_k)$ and a subtraction operation. The computational savings of this reorganization, however, were diminished for larger datasets where memory allocation became as burdensome as calculating convolutions. SCISM was implemented in MATLAB and made use of the TFOCS library [11] to solve the weighted, non-negative LASSO optimization step. Typical analysis ran at a rate of approximately 20 s per profile found.The SCISM source code (written in MATLAB) and its documentation on its usage is available on Bitbucket (https://bitbucket.org/adamshch/scism).

References

1. Cizmár, T. & Dholakia, K. Axial intensity shaping of a Bessel beam. *Proc. SPIE,* 74001Q–74001Q (2009).
2. Hopt, A. & Neher, E. Highly nonlinear photodamage in two-photon fluorescence microscopy. *Biophys. J.* **80,** 2029–2036 (2001).
3. Ji, N., Magee, J. C. & Betzig, E. High-speed, low-photodamage nonlinear imaging using passive pulse splitters. *Nat. Methods* **5,** 197–202 (2008).
4. Harvey, C. D., Coen, P. & Tank, D. W. Choice-specific sequences in parietal cortex during a virtual-navigation decision task. *Nature* **484,** 62–68 (2012).
5. Dombeck, D. A., Harvey, C. D., Tian, L., Looger, L. L. & Tank, D. W. Functional imaging of hippocampal place cells at cellular resolution during virtual navigation. *Nat. Neurosci.* **13,** 1433–1440 (2010).
6. Chaigneau, E., Wright, A. J., Poland, S. P., Girkin, J. M. & Silver, R. A. Impact of wavefront distortion and scattering on 2-photon microscopy in mammalian brain tissue. *Opt. Express* **19,** 22755–22774 (2011).
7. Stirman, J. N., Smith, I. T., Kudenov, M. W. & Smith, S. L. Wide field-of-view, multi-region, two-photon imaging of neuronal activity in the mammalian brain. *Nat. Biotechnol.* **34,** 857–862 (2016).
8. Pati, Y., Rezaiifar, R. & Krishnaprasad, P. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. *Asilomar Conf. Signals Syst. Comput.* 40–44 (1993).
9. Swirszcz, G., Abe, N. & Lozano, A. Grouped orthogonal matching pursuit for variable selection and prediction. *Adv. in Neural Inf. Process. Syst.* 1150–1158 (2009).
10. Apthorpe, N. J. *et al.* Automatic Neuron Detection in Calcium Imaging Data Using Convolutional Networks. Preprint at https://arxiv.org/abs/1606.07372 (2016).
11. Becker, S., Candes, E. & Grant, M. TFOCS: flexible first-order methods for rank minimization. *SIAM Conf. Optim.* (2011).