Supplemental Document

OPTICA

Deep-learning-augmented computational miniature mesoscope: supplement

YUJIA XUE,[1,†] QIANWAN YANG,[1,†] GUORONG HU,[1] KEHAN GUO,[1] AND LEI TIAN[1,2,3,*]

[1] *Department of Electrical and Computer Engineering, Boston University, Boston, Massachusetts 02215, USA*
[2] *Department of Biomedical Engineering, Boston University, Boston, Massachusetts 02215, USA*
[3] *Neurophotonics Center, Boston University, Boston, Massachusetts 02215, USA*
[*] *Corresponding author: leitian@bu.edu*
[†] *These authors contributed equally to this paper.*

Supplement DOI: https://doi.org/10.6084/m9.figshare.20425146

Parent Article DOI: https://doi.org/10.1364/OPTICA.464700

Deep learning-augmented Computational Miniature Mesoscope: supplementary material

This document provides supplementary information for the manuscript "Deep learning-augmented Computational Miniature Mesoscope". First, we provide more details on the CM^2 V2 hardware design, prototype assembly and lateral resolution characterization. Next, we further explain the detailed computational pipeline used in CM^2 V2, including the sparse PSF calibration process, the3D-LSV forward model, and the implementation details of CM^{2}Net. In addition, we elucidate on the distinct role of each network module in CM^{2}Net by performing a series of ablation studies, as well as by visualizing the feature maps from the sub-networks. Third, we provide additional quantitative analysis for the experimental results. Finally, we provide simulation studies on applying the trained CM^{2}Net to imaging complex neuronal structures.

1. CM^2 V2 HARDWARE DESIGN AND PROTOTYPE

A. Additional Details on CM^2 V2 Hardware Design

The CM^2 V2 platform consists of two main parts, including the imaging module and the illumination module. The overall design is visualized in Figs 1a and 2a in the main text. Built from 3D-printed components and off-the-shelf optics and electronics, the CM^2 V2 platform has an overall dimension of 36 mm $\times$ 36 mm $\times$ 15 mm including the backside-illuminated (BSI) CMOS PCB board (IDS Imaging, monochrome BSI CMOS IMX178LLJ, 2076 $\times$ 3088 pixels, 2.4-μm pixel size, 12-bit dynamic range, 58 fps). The entire assembled CM^2 V2 prototype weighs 11.5 grams, of which the CMOS PCB takes most of the size (36 mm $\times$ 36 mm $\times$ 4 mm) and the weight (9 grams) whereas the custom parts weigh only 2.5 grams in-total.

In the imaging module, we use an off-the-shelf rectangular MLA with a 100% fill factor (Fresnel Technologies Inc., no. 630, focal length = 3.3 mm, lens pitch = 1 mm, thickness = 3.3 mm). The MLA is directly placed on top of the CMOS sensor to achieve single-shot 3D imaging with a compact form factor. The MLA is diced into a 3 $\times$ 3 array using a high-precision automatic dicing saw (Disco Dad, no.3220) to keep the sides of the MLA clear in order to reduce vignetting and edge effect. Fig. S1a shows the top and side views of the 3D printed MLA housing and how it is mounted onto the CMOS PCB board (IDS Imaging, board-level IMX178). The red region indicates the sensor region of the CMOS. The green regions are the PCB that host extra electronics.

The hybrid emission filter set consists of two filters: a glass interference filter (Chroma Technology, no. 535/50, 1 mm thickness) and a thin-film absorption filter (Edmund Optics, Wratten color filter no. 12, 0.1 mm thickness). We place the interference filter in front of the MLA and the thin-film absorption filter behind the MLA. This design reduces the maximum angle of incidence onto the interference filter to achieve better filtering whereas the remaining leakage is suppressed by the absorption filter. All the optical elements and the CMOS sensor are held by a 3D-printed housing (designed in TinkerCAD, printed on Formlabs Form 2, black resin, no. RS-F2-GPBK-04). The assembled imaging module is mounted to the CMOS PCB by four mini set screws and hex nuts (Thorlabs, no. HW-KIT3). The assembled CM^2 V2 prototype has a calibrated de-magnification of 1.7$\times$ which results in an effective pixel size of 4.15 μm. The CM^2 V2 prototype does not require precise alignment of the optics. The field varying PSFs are experimentally calibrated and later computationally analyzed by our 3D-LSV model.

The illumination module consists of four freeform LED-illumination units held together by a 3D-printed dome-shaped base plate, which also blocks the ambient light. Each freeform LED-illumination unit contains the following parts: a surface mounted LED (Lumileds, LXML-PB01-0040), a 3D-printed freeform collimator (printed on Formlabs Form 2, clear resin, no. RS-F2-GPCL-04), a glass excitation filter (Chroma Technology, no. 470/40, 4 mm), and a 3D-printed LED housing (printed on Formlabs Form 2, black resin, no. RS-F2-GPBK-04). To achieve better surface quality, we post-process the 3D-printed collimator by the following procedure. A thin layer of clear resin (diluted 5 times with isopropyl alcohol) is applied to the outer parabolic surface of the collimator and cured under a UV lamp. The four illumination units are wired sequentially and connected to an LED driver (LED-dynamics Inc., 3021-D-E-350, 350 mA). In Fig. S1b, we provide

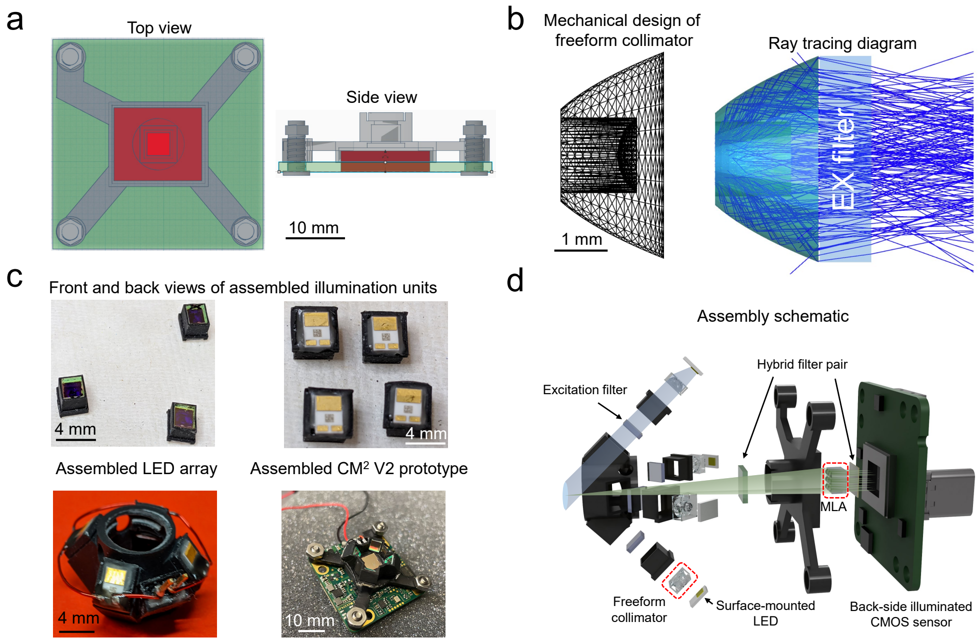

Fig. S1. Additional details on the CM^2 V2 design and the prototype. (a) Top and side views of the 3D printed housing (in gray) that holds the optics (not drawn) and the sensor (in red and green). (b) Mechanical design of the freeform collimator and the ray tracing diagram with the interference-excitation filter from Zemax. (c) Photos of the assembled illumination units, the 4-LED array, and the CM^2 V2 prototype. Each illumination unit hosts a surface mounted LED source, the 3D printed freeform illuminator, and the excitation filter. (d) The complete CM^2 V2 assembly schematic.

the exact mechanical design and the ray tracing diagram of the freeform illuminator. The ray tracing in Zemax has incorporated the LED source spectrum and the incidence-dependent transmittance profiles of the interference-excitation filter from the manufacturer (Chroma Technology). The ray tracing result shows the freeform illuminator efficiently refracts and reflects light from the LED source and allow them to pass through the interference coating. The small divergence of outcoming light is because the LED source has an emitting area of 1 mm^2.

The four LED units are placed symmetrically around the imaging path with a lateral offset of 6.7 mm from the optical axis and tilted by 45 degrees. The positions and orientations of the four illumination units are modeled and optimized in Zemax to achieve the maximum light delivery efficiency and overall excitation uniformity at the imaging plane. The assembled illuminator units (including 3D printed housing, LED source, freeform collimator, and the excitation filter) are shown in Fig. S1c Top. The four units are installed into the LED base plate. Fig. S1c Bottom two photos show the wired illumination array and fully assembled CM^2 V2 prototype, respectively. For better illustration, the CM^2 V2 assembly schematic is provided in Fig. S1d.

After the CM^2 V2 is fully assembled, the excitation is validated on a green fluorescence calibration slide (Thorlabs, no. FSK2). The total excitation power across the designed 8-mm region is up-to 75 mW (at maximum driving current of 350 mA), measured by a power meter (Thorlabs no.PM121D). The illuminator is turned on at 300 mA continuously for 1 hour and no overheating issues were observed.

B. Spectral Properties and the Design of Hybrid Emission Filter Set

The non-normal incidence of uncollimated light in CM^2 V1 results in a substantial wavelength shift of the transmitted spectrum of the emission filter. In Fig. S2a, we plot the spectral profiles of the LED source (blue solid region), green fluorophore emission (green solid region), excitation filter (red curve), emission filter at 0, 20, 40, 60-degree incidence (yellow, purple, green, and blue curve, respectively), and the absorption filter (black curve). The transmittance window of

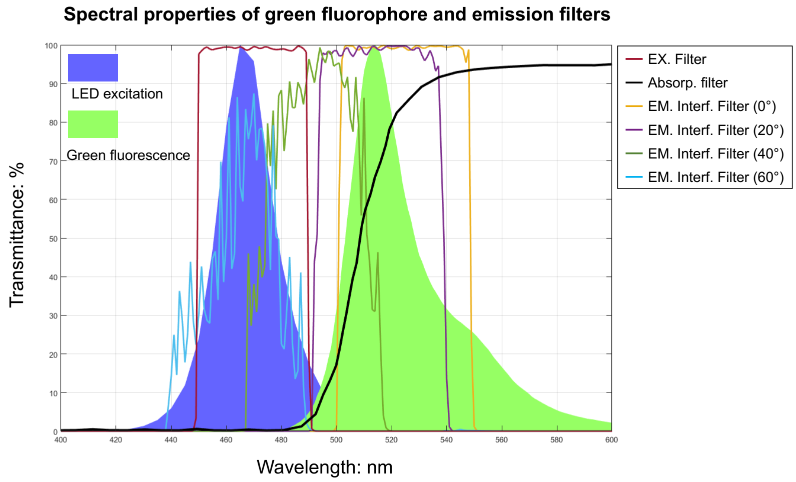

Fig. S2. Spectral properties of hybrid filter design for green fluorescence. The transmission profiles of the interference-emission filter at 0,20,40,60-degree incidence and the incidence-independent absorption filter.

the interference-based emission filter shifts to shorter wavelength with a degraded profile (the oscillating curves) at non-normal incidence. It has a wide spectral overlay with the excitation spectrum, results in a background fluorescence in CM^2 V1 measurements. By adding an incidence-independent absorption filter, whose long-pass spectral profile is plotted in Fig. S2a (black curve). Experimentally, we find that adding the absorption filter to the imaging path provides better suppression of background fluorescence at the cost of slightly reduced transmittance at emission wavelength. Fig. 2e in the main article has clearly shown that the hybrid emission filter provides much enhanced image contrast and reduced the background haziness by $\sim$5 times.

C. CM^2 V2 Lateral Resolution Characterization

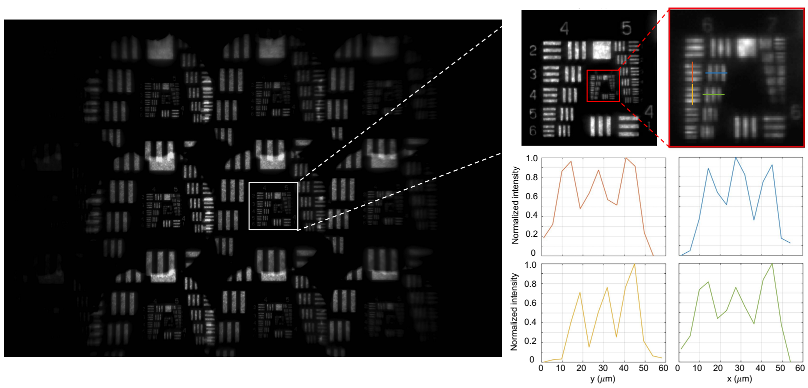

Fig. S3. CM^2 V2 lateral resolution experimental characterization. The experimental measurement on a 2D planar fluorescent resolution target. Zoom-in image show features at both Group 6 Element 3 (6.2-μm) and Group 6 Element 4 (5.52-μm) can be resolved. Plots show the corresponding horizontal and vertical line profiles of the selected elements on the resolution target.

We experimentally characterize the lateral resolution of CM^2 V2 by imaging a fluorescent resolution target. Fig. S3 shows the captured raw measurement of the fluorescent resolution target. The zoom-in region shows that the features at both Group 6 Element 3 (6.2 μm) and Group 6 Element 4 (5.52 μm) can be resolved. For better illustration, we further plot the horizontal and vertical line profiles of the selected elements on the resolution target. Based on this measurement, we conclude that the lateral resolution for CM^2 V2 is $\sim$6 μm.

D. CM^2 V2 Experiment setup

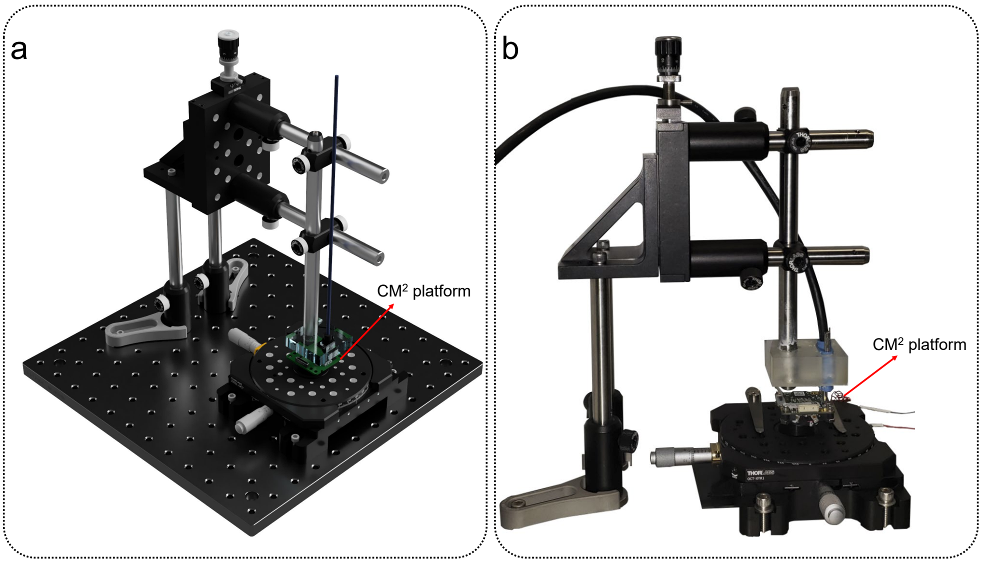

Fig. S4. CM^2 V2 mounting stage. (a) 3D model of the experimental setup. It consists of an XY translation stage with a rotating platform and a single-axis translation stage. A USB Type-C cable connects the sensor with the host computer for data transmission. The four LED units are powered by our custom constant current source, which can output a maximum current of 500 mA (not drawn). (b) Photograph of the mounting stage.

In order to acquire the experimental data, we mount the CM^2 V2 onto the setup shown in Fig. S4. The position of the sample is adjusted by the XY translation stage, and the Z translation stage controls the axial focus of the CM^2 V2. To make sure the measurement is within the range of the calibrated PSFs, a stack of images are collected at different working distances, and the one image with the best reconstruction quality is chosen. The camera parameters are slightly sample-dependent and are generally set to be with 30 ms exposure time, 0 gain, 12- bit dynamic range, and 30 FPS to satisfy our targeted imaging requirements.

2. CM^2 V2 COMPUTATIONAL PIPELINE

A. Sparse PSF calibration process

The spatially varying PSFs of the assembled CM^2 V2 are experimentally calibrated. The calibration process collects the 3D-LSV PSFs at a set of 3D sparse locations within the targeted imaging volume. To perform the calibration, we first build a point source setup consisting of two parts: a point source and a 3-axis high-precision automatic scanning stage. The point source is built by first stacking multiple layers of thin diffusing films (Parafilm) on top of a bright green LED source (Thorlabs, M530L4), whose spectrum approximately matches with the green fluorescence. A 5-µm mounted pinhole (Thorlabs, P5D, stainless steel) is placed immediately after the diffusing films. The multi-layer diffusing film homogenizes the LED illumination before entering the pinhole. The resulting point source provides a large divergence angle (>60 degrees), i.e. a large illumination NA, which is approximated as a point source. The angular intensity distribution from this point source is characterized in Fig. S12b. The point source is mounted on a 3-axis scanning stage (Thorlabs, MT3Z8) that is automatically controlled by a custom MATLAB script.

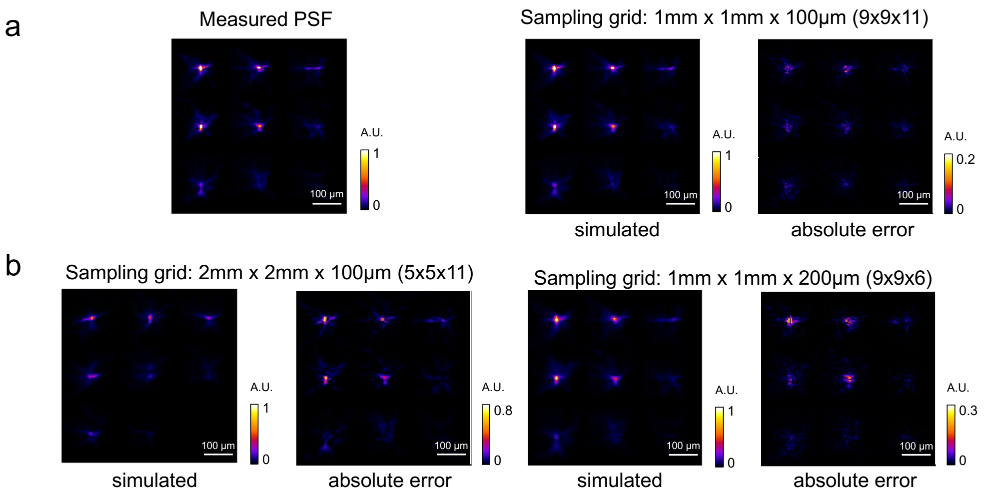

Fig. S5. 3D-LSV model reconstruction accuracy with different PSFs sampling grid. (a) Left: Measured PSF at an unseen location. Right: Simulated PSF and the absolute error using the 3D-LSV model using the current sampling grid. (b) Simulated PSF and the absolute error with 2× down-sampled grid in the xy plane (Left) and 2× down-sampled grid in the axial (z) direction (Right). The absolute error shows the reconstruction accuracy degrades when the sampling grid becomes sparse.

To calibrate the 3D LSV PSFs, we empirically choose a sparse sampling grid and scan the point source across the grid. The 3D grid spans [-4 mm to 4 mm] along both lateral dimensions and [-500 µm to 500 µm] along the axial dimension. The scanning step size is 1 mm laterally and 100 µm axially, respectively. In total, we collect PSFs on a $9 \times 9 \times 11$ grid and 891 measurements.

The sampling grid is essential for the 3D-LSV model to accurately characterize the system. The grid should be dense enough to capture the key variations present in the 3D PSF across the 3D volume while as sparse as possible to allow efficient physical calibration, data storage, and computation. To demonstrate the effect of the sampling grid, we compare the reconstruction accuracy of the 3D-LSV model using the current PSF sampling grid and that from a down-sampled grid in Fig. S5. By comparing the reconstructed PSF at an "unseen" 3D location, the result shows that the reconstruction accuracy degrades when the sampling grid becomes sparse. Specifically, the reconstruction performance with laterally (xy) down-sampled PSFs suffers more severe loss of accuracy compared to that from axially (z) down-sampled PSFs. This is because the PSF variation is more prominent in the xy plane due to the severe off-axis aberrations and PSF truncation at the peripheral FOV. We can further increase the 3D-LSV model accuracy at the cost of PSF calibration time and computational time and memory storage.

After determining the PSF sampling grid, the MATLAB script controls both the stage scanning and provides the triggers for the CM^2 V2 image acquisition. To further account for the nonuniform angular profile of the point source, the MATLAB script adaptively adjusts the exposure times at different scanning locations. This ensures all the measured PSFs are not saturated or too dark. The exposure time is recorded to later normalize the PSF measurements before the 3D-LSV modeling. The whole calibration process takes about 2 hours to complete.

B. Additional details on the 3D Linear Shift Variant (LSV) model

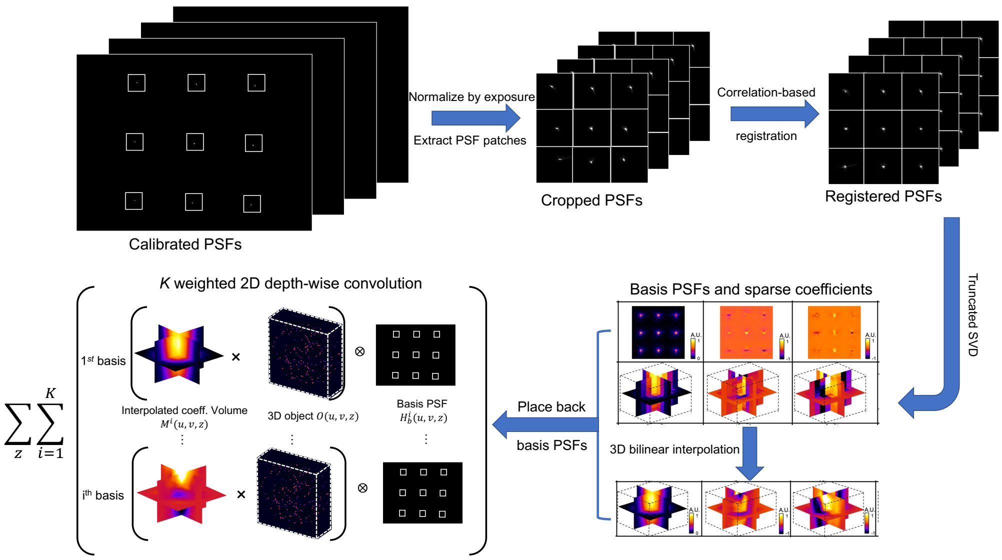

Fig. S6. Procedure diagram of our 3D-LSV model. Local small PSF patches are first cropped from the calibrated PSF measurements based on the chief ray locations. The cropped PSF patches are registered based on the estimated axial sheering and re-grouped into array PSFs. Then we perform TSVD on the registered array PSFs to obtain basis array PSFs and their coefficients at a sparse grid of calibrated locations. The coefficients are then 3D interpolated to the entire imaging volume with 3D bilinear interpolation. Next, the basis PSFs patches are placed back to their original locations. Lastly, the CM^2 measurement is computed by k weighted 2D convolutions between the object volume and the basis PSFs, followed by a summation along the axial dimension z.

We provide additional details of the 3D-LSV model and experimentally validate the accuracy of the model. Fig. S6 shows a diagram of the 3D-LSV decomposition and interpolation process. The experimentally calibrated PSFs (2076 $\times$ 3088 pixels) are first normalized by the exposure time recorded in the calibration process. Next, small PSF patches (160 $\times$ 160 pixels) are extracted from the PSF measurements based on the chief ray locations. Due to the finite conjugate imaging geometry of the CM^2 system, the PSF at a different depth exhibits a different amount of lateral shift that is approximately linearly increases with the depth (i.e. axial sheering). To enable efficient PSF decomposition, we estimate the amount of axial sheering based on the on-axis PSF stack and then align the PSF patches with the in-focus PSF by numerically "undo" lateral shift. The 9 aligned PSF patches are regrouped into a 3 $\times$ 3 foci array (480 $\times$ 480 pixels). Next, the 891 calibrated array PSFs are decomposed using the singular value decomposition (SVD) and truncated to the leading 64 terms. There are two products from this TSVD process: the 64 basis PSFs H (480 $\times$ 480 pixels) and 64 coefficient volumes M (9 $\times$ 9 $\times$ 11 voxels). To match our reconstruction sampling, the coefficient volumes are further interpolated onto a dense 1920 $\times$ 1920 $\times$ 80 grid using the 3D bilinear interpolation method. To construct the full-sized (2076 $\times$ 3088 pixels) CM^2 measurement, the basis PSF patches are then put back to their original pixel locations. The locations are determined by the chief ray locations and the estimated axial sheering. Lastly, the measurement is computed by k weighted depth-wise 2D convolutions between the object volume and the basis PSFs, followed by a summation along the axial dimension z. In Fig. 3c of the main article, we visualize the first 5 basis PSFs and their coefficient volumes. Fig. S7 shows the next 10 basis PSFs and their interpolated coefficient volumes.

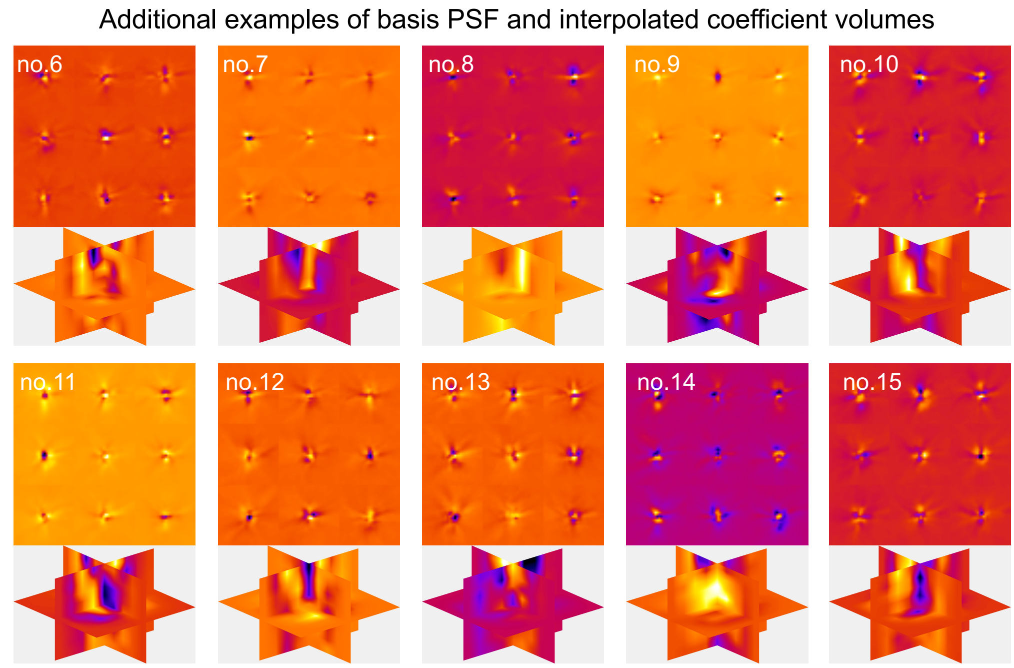

Fig. S7. More examples of the decomposed basis PSFs and the interpolated coefficient volumes.

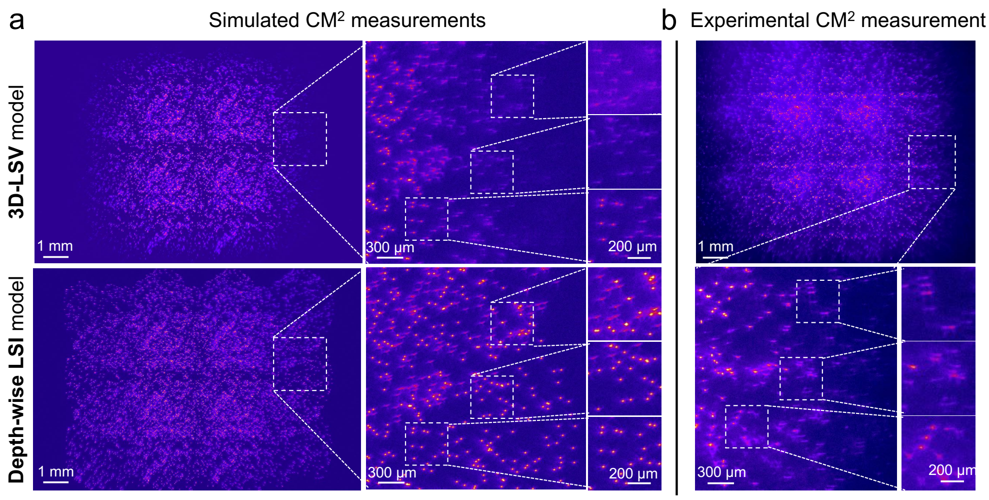

Fig. S8. Experimental validation of the 3D-LSV model. (a) Simulated measurements using our 3D-LSV model and the depth-wise LSI model (sample: 10-μm fluorescence beads randomly placed in a $5 \times 5 \times 0.8$ mm^3 volume). (b) Experimental measurement on 10-μm fluorescence beads with similar density to the simulation. This comparison shows that our 3D-LSV model can accurately capture the key aberrated features in real experiments.

To visually validate the synthetic measurement generated by this 3D-LSV model, we compare a simulated image and an experimental measurement taken from a $5 \times 5 \times 0.8$ mm^3 volume with 10-μm fluorescence beads with similar seeding density in Fig. S8. The forward model used in our V1 system is based on a depth-wise linear shift invariant (LSI) approximation[1]. To highlight the improvement of the proposed 3D-LSV model over the depth-wise LSI model, we compare the synthetic measurements from the same object using both models in Fig. S8a. In the zoom-in panels taken from the same regions at the peripherical FOV where the aberrations are more apparent, we highlight that the 3D-LSV model can synthesize the key image features in the real experiment. This is essentially important to train CM^{2}Net based on only simulation data, yet the trained network is directly generalizable to real experimental measurements.

C. Analysis on the denoising capability of the 3D-LSV model

In this section, we analyze the denoising capability of the 3D-LSV model with different number of basis PSFs. We first decompose the raw calibration PSF stacks with different number of basis using the truncated Singular Value Decomposition (TSVD) and then compute the corresponding 3D-LSV model and simulate the PSF at a seen position. Next, we compare the absolute error between the simulated and the measured PSFs and plot the background intensity profile. Since when constructing our 3D-LSV model only the simulated PSF is used even at the calibration "seen" 3D locations, this procedure allows assessing the effective denoising capability of the 3D-LSV model. Fig. S9 shows that all simulated PSFs have smoother background profiles, validating that the low-rank model can suppress the noise particular in the background region. Moreover, the denoising performance for 3D-LSV model is related to the number of basis. When the number of basis is too large, the LSV model not only well approximates the system aberrations, but also inadvertently capture the random noise present in the raw calibration PSF measurements. This shows a trade-off between the reconstruction accuracy and denoising capability. Since the noise level at for our raw PSF measurement is low, we choose a relatively large number of basis PSFs (K = 64) to provide a good reconstruction accuracy with a moderate denoising effect.

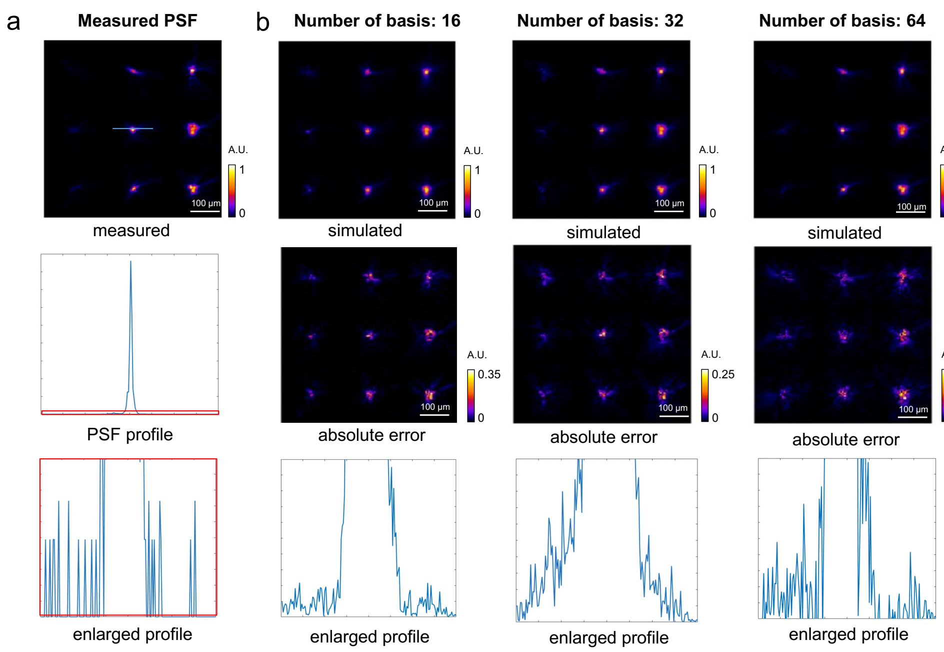

Fig. S9. Denoising capability of the 3D-LSV model with different number of basis PSFs. (a) Top: A measured 3×3 PSF array on the calibration grid. Middle: Intensity profile of the labeled PSF. Bottom: Enlarged PSF intensity profile to show the background fluctuations. (b) The simulated PSF with different number of basis PSFs, including 16, 32, and 64 (used in the main text). The absolute error between the simulated and measured PSF shows that the 3D-LSV model accuracy improves when the number of basis PSFs becomes larger. The enlarged profiles for all simulated PSFs show smoother background intensity profiles compared to the measured PSF, which indicates that the low-rank model can suppress the noise in the raw PSF measurement.

D. Implementation details of the CM^{2}Net and its building block

In this section, we describe the details of the network implementation and its training and testing procedure. The CM^{2}Net contains three sub-networks, including the view demixing-net, view-synthesis-net, and lightfield-refocusing enhancement-net. The three sub-networks all use the same residual network structure and only differ in the input/output dimensions. The detailed structure of a residual block is provided in Fig. S10. The input tensor first goes through a 2D convolution layer which has 64 convolution kernels with size 3 × 3. After the 2D convolution layer, the intermediate tensor is fed into a batch normalization layer, followed by a nonlinear activation layer using Parametric ReLU (PReLU) layer, a nonlinear function with learnable slope in the negative side of the axis. The tensor then goes through another pair of 2D convolution and batch normalization layers to further increase the receptive field. Lastly, the tensor is element-wise added to the input tensor of this block, which forms a residual connection. The numbers of input and output channels, denoted by (a, b) for the demixing-net, view-synthesis-net, and enhancement-net are (9, 9), (9, 80), and (32, 80), respectively.

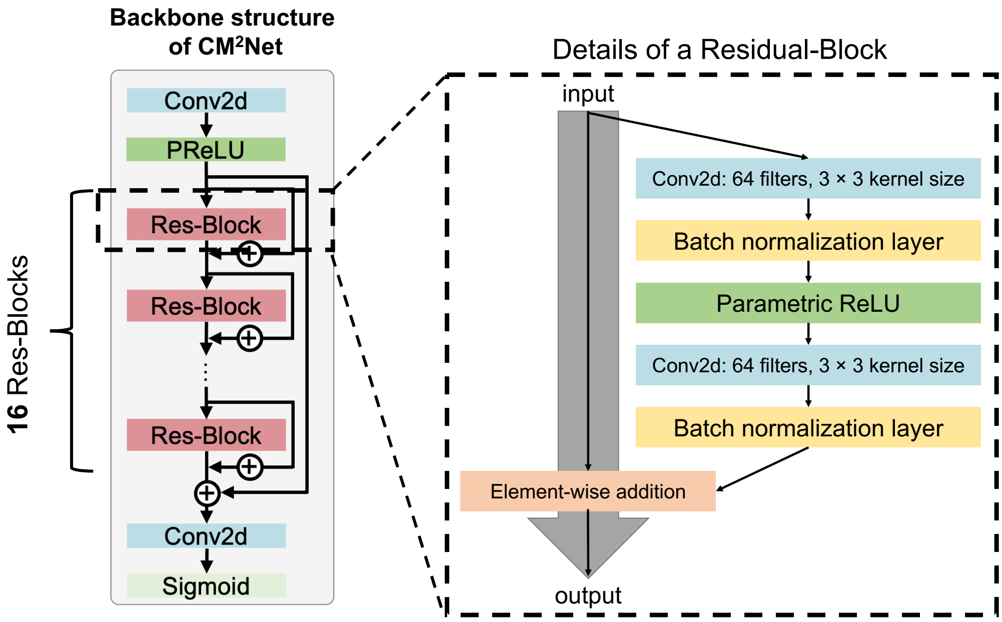

Fig. S10. Implementation details of CM^{2}Net building block. The three sub-networks in the CM^{2}Net share the same backbone structure of a deep residual network that connects 16 residual-blocks with sequential residual connections. Each residual-block consists of a 2D conv layer with batch normalization (BN), a Parametric ReLU nonlinearity, another 2D conv layer with BN, and element-wise addition.

The CM^{2}Net takes a stack of 9 cropped views from a CM^2 measurement as the input. In training phase, the input data dimension to demixing-net is $320 \times 320 \times 9$, where the third dimension denotes the number of channels. The output size from the demixing-net remains the same. The non-learnable lightfield refocusing module applies a "shift-and-add" refocusing operation on the demixed 9 views:

$$RFV(x, y; \Delta z) = \sum_{u=-1}^{1} \sum_{v=-1}^{1} VS\left(u, v; x - \frac{M^2 d}{z_0} u\Delta z, y - \frac{M^2 d}{z_0} v\Delta z\right), \tag{S1}$$

where RFV is the refocused slice with Δz defocus distance, (u, v) is the coordinates of the 3×3 microlens array, VS is the demixed view stacks, the amount of shift is determined by the system magnification M, nominal focal distance z_0, size of the microlens d, the microlens index (u, v) and the defocus distance Δz. To implement this, we convert the amount of shift to be [-18: +17] pixels (+ means shifting towards the center view, - means the opposite), which ensure that the refocused stack consisting of 36 planes will cover the targeted 800-µm depth range. Note that this refocusing algorithm generates artifacts around the image boundaries. Therefore, we discard the outmost 32 pixels in both lateral dimensions, resulting in a $256 \times 256 \times 36$ refocused volume.

The enhancement-net is trained to first resample the refocused volume onto an axially finer sampling grid ($256 \times 256 \times 80$) matching the ground-truth volume, and then enhance the 3D reconstruction. The resampling is performed by a 2D convolution layer of kernel size 3 with 80 channels, which increases the depths from 36 to 80. To perform the view-synthesis, similarly only the central $256 \times 256 \times 9$ region is extracted from the demixed views and fed into the view-synthesis-net. The output dimension from the view-synthesis-net is $256 \times 256 \times 80$, the same as the ground-truth volume.

In the inference phase, we directly feed the full-FOV measurement to perform a single-pass 3D reconstruction, which bypasses the stitching artifacts suffered by patch-wise reconstructions.

E. Ablation study on the demixing module of the CM^{2}Net

In this section, we show in simulation that the view-demixing task can be reliably performed and the demixing-net module substantially improves the quality in the downstream 3D reconstruction.

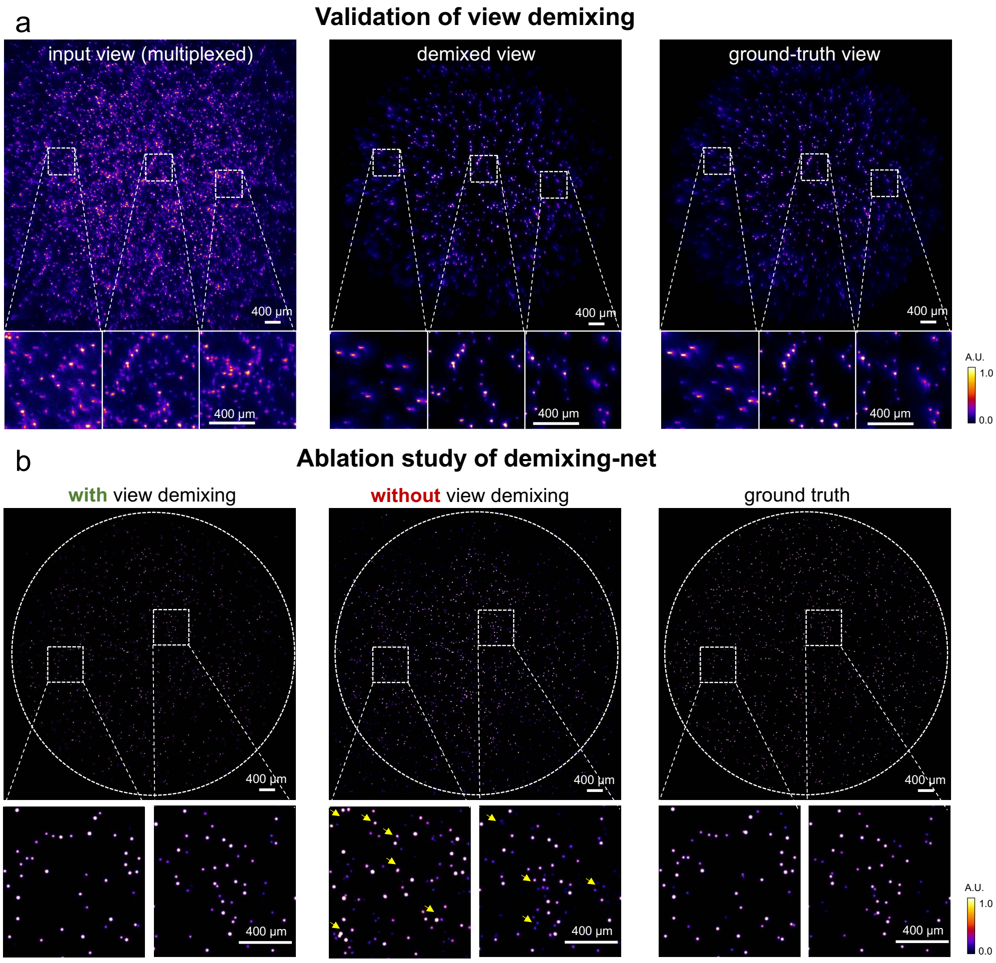

Fig. S11. View-demixing network module significantly improves 3D reconstruction quality.(a) View demixing results on a simulated CM^2 measurement (sample: fluorescent beads with random sizes between 10 μm and 20 μm in a cylindrical volume (7-mm diameter, 0.8 mm depth). The multiplexing artifacts in the "input view" from the raw CM^2 measurements are effectively demultiplexed by the demixing-net. (b) Ablation study results on the demixing-net. The XY MIPs shows that the proposed CM^{2}Net structure with the demixing-net module provides high-quality reconstruction. Removing the demixing-net module results in a large number of false positives (marked by yellow arrows) in the reconstruction.

First, we demonstrate the effectiveness of view-demixing by comparing the demixing-net demixed views against the ground truth in simulation. Both the CM^2 measurement and the nine individual crosstalk-free views are simulated using our 3D-LSV model using the PSFs of the MLA and each microlenses, respectively. In Fig. S11, we show results on a testing volume consisting of fluorescent beads with random sizes between 10 μm and 20 μm in a cylindrical volume (diameter

∼7 mm, thickness ∼0.8 mm). The large FOV results in strong view multiplexing in the raw CM^2 measurement, as evident in the example "input view" centered around the central microlens. The "demixed view" from the demixing-net closely matches with the ground-truth view from the central microlens without crosstalk artifacts, as highlighted in the three zoom-in panels taken from widely separated regions. The demixing-net can robustly perform this task since each fluorescent bead imaged by different microlenses contain distinct aberrated features.

Next, we show that the view-demixing step significantly improves the 3D reconstruction quality. Specifically, we perform the following ablation study in Fig. S11b. The same dual-branch reconstruction network (including the lightfield-refocusing enhancement module and the view-synthesis module) is used to process either the demixing-net demixed views or the multiplexed views from the raw measurement on the same simulated data. The XY maximum intensity projections (MIPs) of the 3D reconstructions from the two networks and the ground-truth volume are shown in Fig. S11b. As highlighted in the two zoom-in regions, the reconstruction from the network with the demixing-net closely matches with the ground truth, whereas the one without view-demixing suffers from a large number of false positives (marked by the yellow arrows).

We further quantify the reconstruction quality by recall and precision. The CM^2Net with the demixing-net achieves an average recall and precision of 0.7 and 0.94 respectively, while the one without the demixing-net achieves much lower recall of 0.57 and precision of 0.24. This comparison highlights that although the CM^2Net without the demixing-net is sensitive enough to recover 57% of the emitters, the 3D reconstruction suffers from very low precision. The comparison on the precision values implies that the demixing-net helps to reduce the false-positive rate from 0.76 to 0.06, a 13× improvement.

F. Ablation study on the reconstruction module of the CM^2Net

In this section, we perform an ablation study to analyze the potential reasons for the decrease of recall near the edges of the FOV. We first separate the CM^2Net into demixing network and the reconstruction network (consisting of the enhancement-net and the view-synthesis-net). First, we train another network with only the reconstruction network and take the ground-truth demixed views as the input. The recall for emitters at different lateral location are then evaluated on the testing set using the same method in Section 3.2. The reconstruction results achieve recalls close to 1 consistently for the entire 7-mm FOV range (Fig. S12a yellow curve), showing the effectiveness and robustness of our reconstruction network design. To further evaluate the trained CM^2Net, we quantify the recall of the reconstruction network in the pre-trained CM^2Net by directly inputting the ground-truth demixed views. The result (orange curve) shows a slight degradation as compared to the re-trained reconstruction network but remain >0.89 across the 7-mm FOV. Both results are much higher than that from the CM^2Net predictions on the raw CM^2 measurement (blue curve). This indicates that the CM^2Net's reconstruction module provides superior performance, and the degraded performance at the outer FOV is due to imperfect view-demixing at these regions.

Upon visual inspection, we hypothesize that the view-demixing network is impacted by the low light intensity and reduced SNR of the point source near the edge of the FOV. To show this, we calculate the intensity map of the point source under the central microlens in Fig. S12b and show that the point source's intensity drops as much as ∼85% near the edge as compared to the intensity at the central FOV. Moreover, we compare the point source intensity map and recall map on the same pixel grid. Both maps exhibit similar non-isotropic distributions and significant drops at outer FOV regions. Finally, we show comparisons between the ground-truth and the view-demixing-net predicted demixed views (1-mm^2 patch) for both a central and an edge image patches in Fig. S12c. The view-demixing result for the central FOV is highly accurate, whereas the result near the edges suffer from a large amount of missing particles. We also visually observe that the SNR for the demixed views at the edge FOV are much worse than the central FOV.

Overall, this ablation study shows that the combination of the view-synthesis-net and enhancement-net can achieve superior reconstruction performance and is robust to large SNR variations and high dynamic range of the measurements. We provide evidence that the low recall near the edge is possibly due to the unevenly distributed point source used in the PSF calibration.

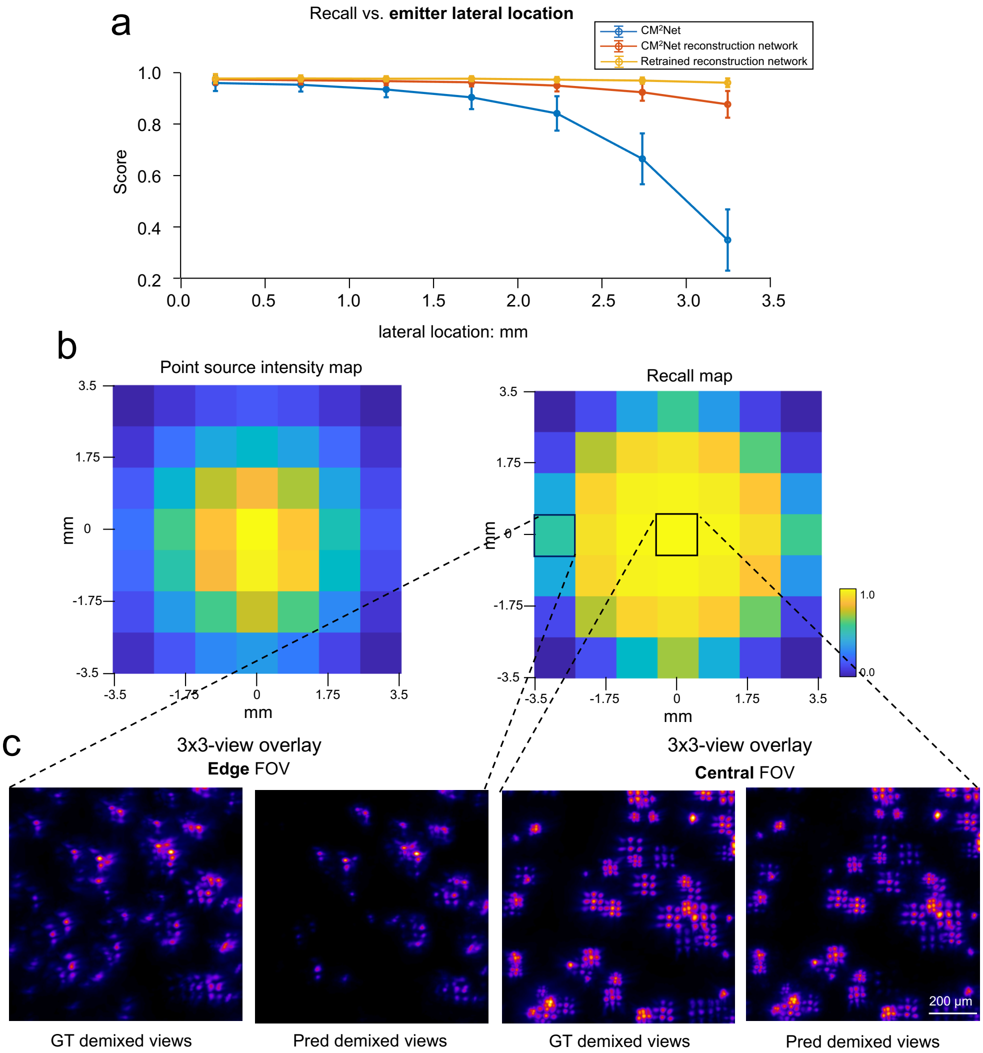

Fig. S12. Ablation study on the reconstruction module of the CM^{2}Net using the ground-truth demixed views.(a) Yellow curve: recall for a retrained reconstruction module using the ground-truth demixed views as the input. The recall is close to 1 consistently for the entire 7-mm FOV range, showing the effectiveness and robustness of our reconstruction module design. Orange curve: recall for directly feeding the ground-truth demixed views to the reconstruction module of the pre-trained CM^{2}Net. The recall remains >0.89 across the 7-mm FOV. Blue curve: recall for the trained CM^{2}Net, which is equivalent to feeding the view-demixing-net demixed views to the reconstruction module of the pre-trained CM^{2}Net. The decreased performance is attributed to the degraded view-demixing results. (b) Left: the intensity distribution of the point source used in our PSF calibration as measured by total intensity under a single microlens at different scanning positions. Right: the recall map of the CM^{2}Net is visualized by binning onto the same pixel grid as the point source intensity map. Both maps exhibit much reduced values at the outer FOV regions, which suggests that the degraded recall at the outer FOV regions may be due to the rapid intensity decay of the point source used in the model. (c) Comparison of the 3×3-view overlay view-demixing results between patches from the edge and central FOV. The central patch that achieves $\sim$1 recall has higher SNR and shows nearly perfect view-demixing result. However, the view-demixing prediction on the patch from the edge FOV suffers from low SNR and missing particles.

G. Ablation study on the reconstruction sub-modules of the CM^{2}Net

In this section, we perform ablation studies on the two branches in the reconstruction module to show how the light-field refocusing (LFR) enhancement branch and the view-synthesis (VS) branch complement each other.

First, we compare the network performance by removing either the LFR or the VS branch in the CM^{2}Net to show the effectiveness of the combined network structure. To do this, we build two new reconstruction networks with only the LFR or the VS branch (the network structure is shown in Fig. S13a). Both networks use the same data set and follow the same training scheme as the full CM^{2}Net. After training for 48 hours on all the networks, we test and evaluate the performance by the quantitative analysis procedure shown in Section 3.2. The results are summarized in Fig. S13b. This ablation study shows that the CM^{2}Net achieves the best F1-score on all the conditions by combining both the VS and LFR branches.

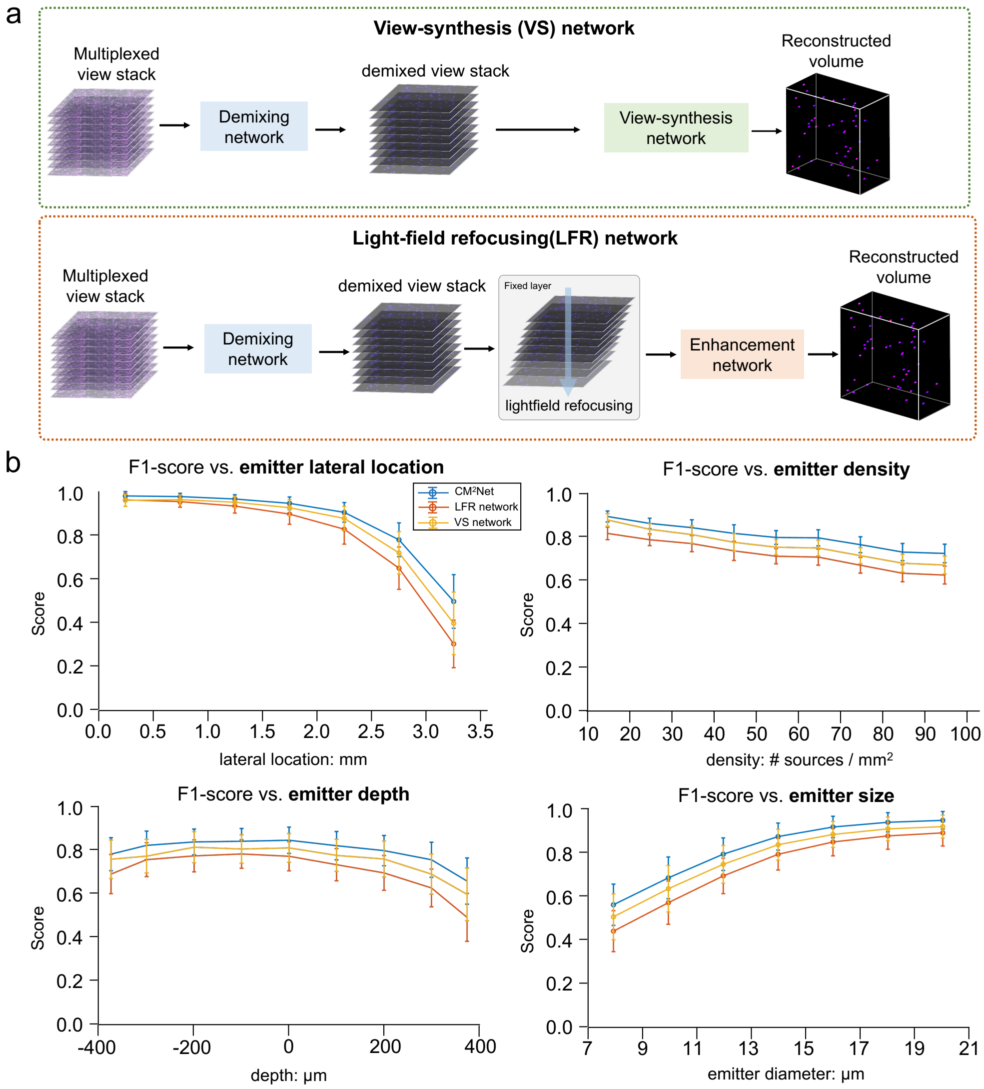

Fig. S13. Ablation study on CM^{2}Net's reconstruction sub-nets. (a) The network structure to perform ablation study on the view-synthesis branch (VS) and the light-field refocusing enhancement branch (LFR). (b) Quantitative analysis of evaluating the particle detection performance of the VS, LFR and CM^{2}Net. CM^{2}Net achieves the best score at all conditions by combining the complementary information from the VS and LFR branches.

To further illustrate the LFR and VS networks provide complementary information. Next, we

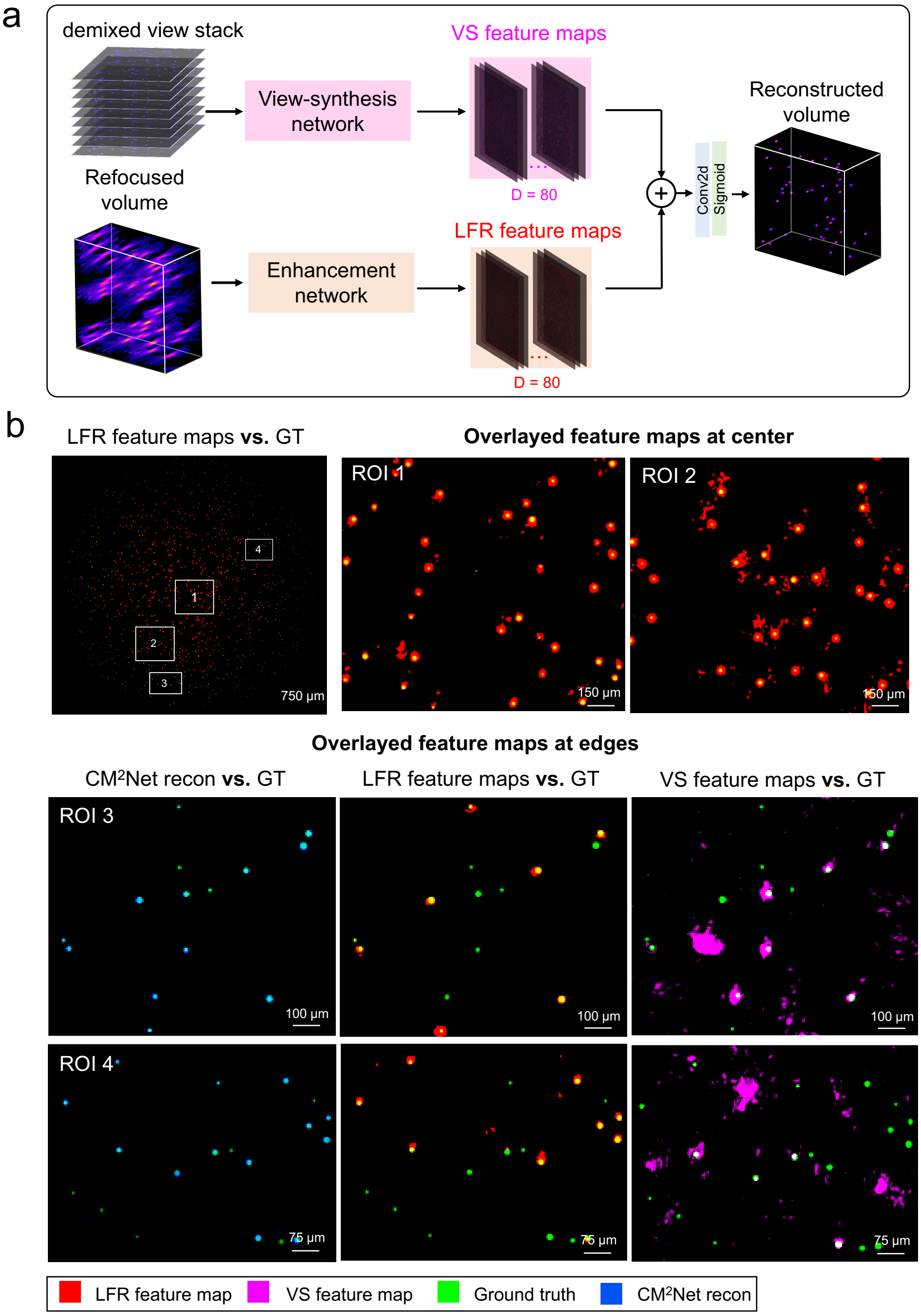

Fig. S14. Feature maps analysis of the CM^{2}Net's reconstruction sub-nets. (a) Illustration of the extraction of feature maps from the VS and LFR branches. (b) Top: The overlay between the MIPs of the ground truth and the feature maps from the LFR network. The contrast for feature maps is enhanced for visualization. The ROI 1 and ROI 2 show that the LFR feature maps highly matches with the ground truth at the central FOV (yellow indicates matched particles). Bottom: The overlay of the MIPs of the ground truth and the CM^{2}Net reconstruction, LFR feature maps MIP and VS feature maps MIP at the FOV edges. The particles detected by the CM^{2}Net (in light blue) is the combination of the particle detected by the LFR branch (in yellow) and the VS branch (in white). The CM^{2}Net learns the complementary information provided by the VS and LFR branches to achieve high detection rate across a wide FOV.

visualize the final feature maps from the two branches of the CM^{2}Net. We directly extract the feature maps from the trained CM^{2}Net (as shown in Fig. S14a) and compare the MIP of the feature maps with the MIP of the ground truth at the central and edge FOVs. In Fig. S14b, we show that the feature maps from the LFR branch highly matches with the ground truth at the central FOV while miss a large amount of particles at the edge. This is because the input (lightfield refocused volume based on the demixed views) to the LFR-network suffers from artifacts at the peripheral regions from inexact view matching from the demixed views and the boundary artifacts from the "shift-and-add" operation. To address this challenge, CM^{2}Net uses the additional VS branch to the reconstruction module to help improve the performance at the peripheral FOV regions. In Fig. S14b, we compare the ground truth volume, the CM^{2}Net reconstruction, the feature maps from the LFR branch, and the feature maps from the VS branch to visualize the improvement owing to VS branch. We show that at the FOV edges, the VS branch can detect particles that are missed by the LFR branch, and thus improve the CM^{2}Net's recovery of emitters across the large FOV.

3. PERFORMANCE EVALUATION METRICS

We use particle-wise detection metrics, including recall, precision and F1-score to quantitatively evaluate the CM^{2}Net 3D reconstruction. We choose this set of metrics because they focus on quantifying the error made on the reconstructed emitters, whereas voxel-value averaged metrics, e.g. mean squared error, are biased by the background for sparse objects like ours.

We describe the details on the computations of the three metrics using MATLAB language. Starting from the reconstructed volumes from the CM^{2}Net, we first detect all the recovered particles within the volume. The detection is done by binarizing the recovered 3D volume with a global optimal thresholding (`imbinarize`) and then extracting the locations and sizes of connected 3D components (recovered particles) from the binarized volume (`bwconncomp`, `regionprops3`). We then compute a distance matrix by calculating the Euclidean distance between every recovered and ground-truth particle and solve a linear assignment problem based on the distance matrix to assign the recovered particles to the corresponding ground truth (`matchpairs`).

If the Euclidean distance between the recovered and the matched ground-truth particles is larger than a pre-defined distance threshold (i.e. 2× the particle size in our simulation), we count it as a False Negative (FN), meaning that the CM^{2}Net fails to correctly reconstruct the particle. If a recovered particle does not find a match in the ground truth, we count it as a False Positive (FP). A True Positive (TP) is when the distance between the recovered and the matched ground-truth particles is smaller than the distance threshold (examples are shown in Fig. S16a). Lastly, the recall, precision, F1-score, Jaccard index (JI), and lateral and axial Root mean squared localization error (RMSE) are computed as follows:

$$\text{Recall} = \frac{\#\text{TP}}{\#\text{TP} + \#\text{FN}} = \frac{\#\text{emitters correctly reconstructed}}{\#\text{total emitters in GT volume}}, \tag{S2}$$

$$\text{Precision} = \frac{\#\text{TP}}{\#\text{TP} + \#\text{FP}} = \frac{\#\text{emitters correctly reconstructed}}{\#\text{total emitters in reconstructed volume}}, \tag{S3}$$

$$\text{F1 score} = \frac{2}{\text{Recall}^{-1} + \text{Precision}^{-1}} = \frac{\#\text{TP}}{\#\text{TP} + 1/2(\#\text{FP} + \#\text{FN})}, \tag{S4}$$

$$\text{JI} = \frac{\#\text{TP}}{\#\text{TP} + \#\text{FP} + \#\text{FN}}, \tag{S5}$$

$$\text{Lateral RMSE} = \sqrt{\frac{1}{\#\text{TP}} \sum_{i \in R \cap G} (x_i^r - x_i^g)^2 + (y_i^r - y_i^g)^2}, \tag{S6}$$

$$\text{Axial RMSE} = \sqrt{\frac{1}{\#\text{TP}} \sum_{i \in R \cap G} (z_i^r - z_i^g)^2}, \tag{S7}$$

where x_i^r, y_i^r, z_i^r is reconstructed emitter's position and x_i^g, y_i^g, z_i^g is the corresponding ground truth position.

Recall and precision quantify the detection performance in two complementary aspects. Recall quantifies the CM^{2}Net's sensitivity by the fraction of emitters it correctly reconstructed out of the total in the ground-truth volume. Precision measures the fraction of emitters it correctly reconstructed out of the total reconstructed emitters. Both F1-score and JI are metrics combining recall and precision, which are highly correlated. Lateral and axial RMSE measure the localization accuracy. The global threshold for binarizing the recovered volume is set by maximizing the F1-score on the evaluation set [2]. The recall, precision, lateral and axial RMSE under different conditions are reported in the main text (Section 3.2 and Figure 6). The F1-score and JI under the same set of conditions are shown in Fig. S15.

The metric maps reported in Sections 3.2 and 3.3 are computed by the metric value on non-overlapping patches (250 µm × 250 µm in Section 3.2 and 500 µm × 500 µm in Section 3.3). When quantifying the metrics on experimental data (Sections 3.3 and 3.4), we first co-register the CM^{2}Net reconstruction and the widefield measurement by manually selecting a few matched feature points in the central FOV. We then extract particles by binarizing both the widefield measurement and the CM^{2}Net reconstruction. Next, we follow the same evaluation pipeline as above. Since the amount of image distortion suffered by the widefield and CM^2 measurements are different, good registration can only be achieved around the central region, while particle pairs at the peripheral region are expected to have a larger amount of separations. Therefore, the

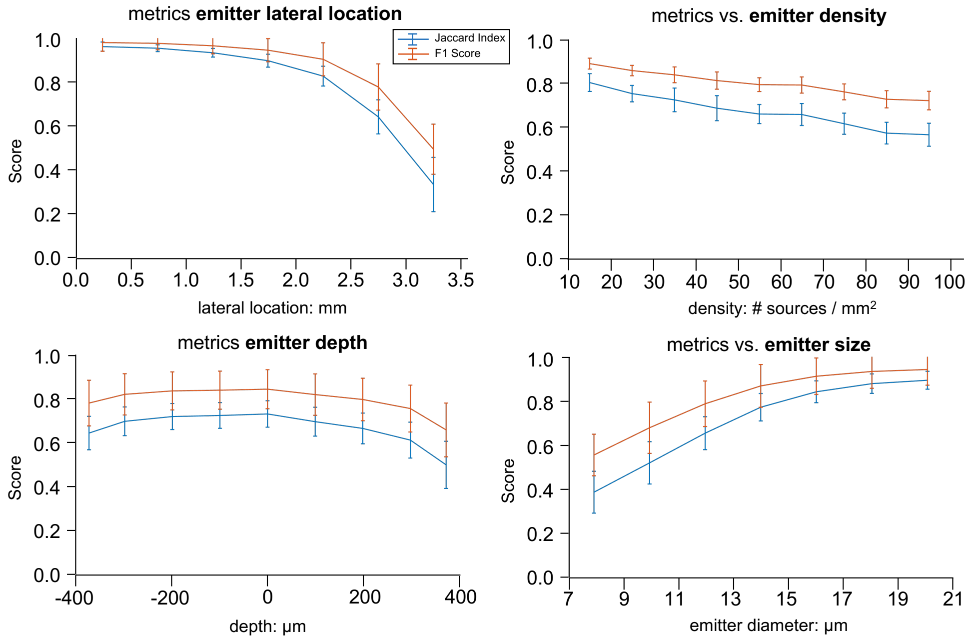

Fig. S15. CM^{2}Net performance evaluated by F1-score and JI. Quantified F1-score and JI when varying the emitter's (a) lateral location, (b) seeding density, (c) depth, and (d) size. F1-score and JI show the same trends since they both combine recall and precision in a similar way.

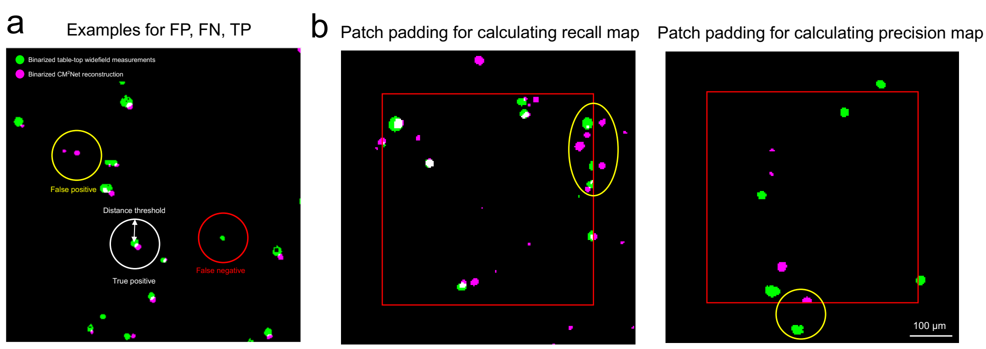

Fig. S16. Details for the procedure of CM^{2}Net quantification analysis.(a) An example overlay image patch between the registered and binarized widefield measurement and the CM^{2}Net reconstruction. The circle indicates the distance threshold for performing linear assignment. The yellow, red, and white circles show examples of false positive, false negative, and true positive, respectively. (b) Example patch padding procedure for computing the recall (left) and precision (right) map, respectively. The red box shows the effective 500 $\times$ 500-μm^2 non-overlap FOV used to quantify the metrics. To maintain the distance threshold at the FOV edge, we choose a larger CM^{2}Net reconstruction patch (700 $\times$ 700-μm^2) when calculating the recall map, and a larger widefield measurement patch (700 $\times$ 700-μm^2) when calculating the precision map. The yellow circle shows the particles around the edges, which requires the patch padding to find a proper match within the defined distance threshold.

distance threshold we choose for the experimental data is $\sim 5\times$ the particle diameter, which is larger than the one used for the simulation ($2\times$ the particle diameter).

To maintain the same distance threshold for particles at the edge of each patch, we perform particle-pairing on slightly expanded patches (see in Fig. S16b). Specifically, when computing the recall map on a 500×500-μm^2 patch, we select a larger CM^{2}Net reconstruction patch (700×700-μm^2) and pad 0 to the corresponding widefield measurement patch, since recall requires detecting all the true positives. We visually show that without considering the distance threshold at the edge, the recall value will decrease since some matched pairs (yellow circle) between the reconstructed and the matching ground-truth are separated by the edge of the patch (red box) while within the distance threshold. When calculating the precision map, we select a larger widefield measurement patch (700×700-μm^2) and pad 0 to the corresponding CM^{2}Net reconstruction patch in order to identify all the reconstructed particles corresponding to the matching widefield measurement. We visually show that without considering the distance threshold at the edge, some matched pairs (yellow circle) are separated by the edge of the patch (red box) while within the distance threshold, which contaminates the precision quantification.

The binarized MIP image pairs are shown in Figs. S20b and S22b for the 10-μm and mixed-size bead phantom, respectively. Moreover, since the particles in the experiment are not uniformly distributed, we observe `NaN` values (both the widefield measurement and the reconstruction in this image patch are empty) when perform quantification for the experimental data. For fair evaluation, we exclude the `NaN` values when calculating the mean and standard deviation. When making the metric maps, we linearly interpolate the pixels having `NaN` values, and applies a binary mask to indicate the sample region.

In Fig. 6, the line plots show the quantitative metrics under different conditions. Each point and the associated error bar represent the mean and standard deviation, respectively. The testing data set contains in-total 180 random volumes and $\sim 5\times 10^5$ emitters. Each point in Fig. 6b is computed on $\sim$20 volumes for each range of emitter density. Each point in Fig. 6a, c and d is computed on emitters for each range (labeled on the horizontal axis) of lateral location, depth, and emitter size.

4. EXPERIMENTAL PREPARATION AND RESULT ANALYSIS

A. Preparation of phantom objects and table-top widefield measurements

Fluorescent 3D objects are prepared according to the following protocol. Green fluorescent particles (Thermo Fisher Scientific, Fluoro-Max Dry Fluorescent Particles, 10-µm and 15-µm) are added to $\sim$1 mL of clear resin (Formlabs, no. RS-F2-GPCL-04). The mixture is then diluted and moved onto a standard 3 $\times$ 1 inch microscope slide (Thermo Fisher Scientific, no. 125493) and curved under a UV lamp for $\sim$30 minutes. We use a micropipette (Thermo Fisher Scientific, Adjustable Volume Pipette, 10-100 µL, no. FBE00100) to control the transferred volume of mixture ($\sim$40 µL) so that the object is $\sim$800 µm thick and 7-mm wide.

To verify the 3D reconstructions from CM^{2}Net, we experimentally collect axial focal stacks (z-stack) on a commercial table-top epi-fluorescence microscope (Nikon TE2000-U) with GFP filter sets (Thorlabs, no. MDF-GFP) and a scientific CMOS camera (PCO Imaging, Pco.edge 5.5). To acquire the full-FOV widefield measurements, we use a low-magnification objective lens (Nikon, CFI Plan Apo Lambda 2$\times$, 0.1 NA) with 50-µm axial step size across the 800-µm depth range. To acquire a high-resolution axial scan at zoom-in regions, we use a high-magnification objective lens (Nikon, CFI Plan Achromat 20$\times$, 0.4 NA) with 25-µm axial step size across the 800-µm depth range.

B. Preprocessing steps on experimental data

In this section, we provide additional details on the preprocessing of experimental CM^2 measurements. The goal is to match the intensity distribution between the simulated and the experimental data so that our simulator-trained CM^{2}net can generalize well to real measurements. The main discrepancy between the measured and simulated images is the low-frequency fluorescent background due to the reflection from the microscope slide. To remove the background and match the intensity statistics, we preprocess the experimental measurements with histogram matching. The reference histogram is estimated from the entire training dataset. The histogram matching builds a monotonic mapping from the experimental data to the simulated data. A comparison between the histograms of the simulated data, the experimental data with and without histogram matching are provided in Fig. S17, where one can clearly see the background is much suppressed after the preprocessing. This can be further verified in their histograms. The distribution of preprocessed experimental data matches simulated data much better than raw experimental measurements. We find that other commonly used background removal methods (such as thresholding, morphological opening, etc.) do not generalize well on real experimental measurements. Another preprocess step is to adjust the centers of the cropped 9 views to compensate for the small displacement of the chief ray positions between the simulation and experiments. The adjustment is done by manually aligning a few on-axis particles in each view.

C. Recall, precision, and F1-score maps in phantom experiments

In this section, we report the details of recall, precision, and F1-score evaluations on the two experiments in the main article. The recall and precision maps are computed on non-overlapping patches (500 µm $\times$ 500 µm) across the entire FOV.

For the 10 µm beads phantom in Fig. 7, the recall and precision maps are shown in Fig. S20a. The CM^{2}Net reconstruction achieves a recall $\sim$0.78 and precision $\sim$0.80. In comparison, the recall and precision in simulation at the corresponding density range is $\sim$0.83 and $\sim$0.97, respectively (Fig. 6b). This shows that the simulator-trained CM^{2}Net degrades slightly on experiments with $\sim$5% higher mis-detection rate and $\sim$7% higher false-positive rate at this imaging condition. To further analyze the potential reason that causes the performance decrease in the experiment, we compare the widefield measurement, the 3×3 overlay of the predicted view-demixing results, and the CM^{2}Net reconstruction MIP. We find that the decreased performance originates from the view-demixing results (see Fig. S18). We attribute the reduced performance to the undesired extra views in the experimental measurements due to an extra column of partial microlenses adjacent to the main 3 $\times$ 3 microlens array (see Fig. S19).

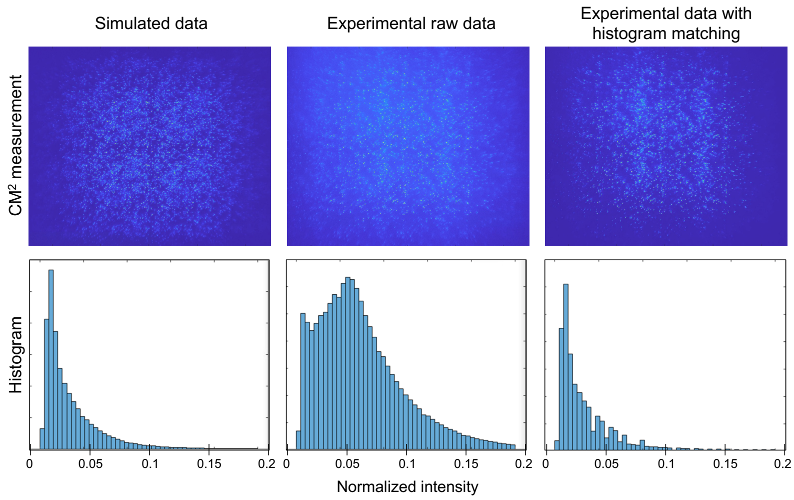

Fig. S17. Preprocessing on experimental CM^2 measurements with histogram matching. Preprocessed experimental measurements have a suppressed background and a better match with the intensity distribution of the simulated training data.

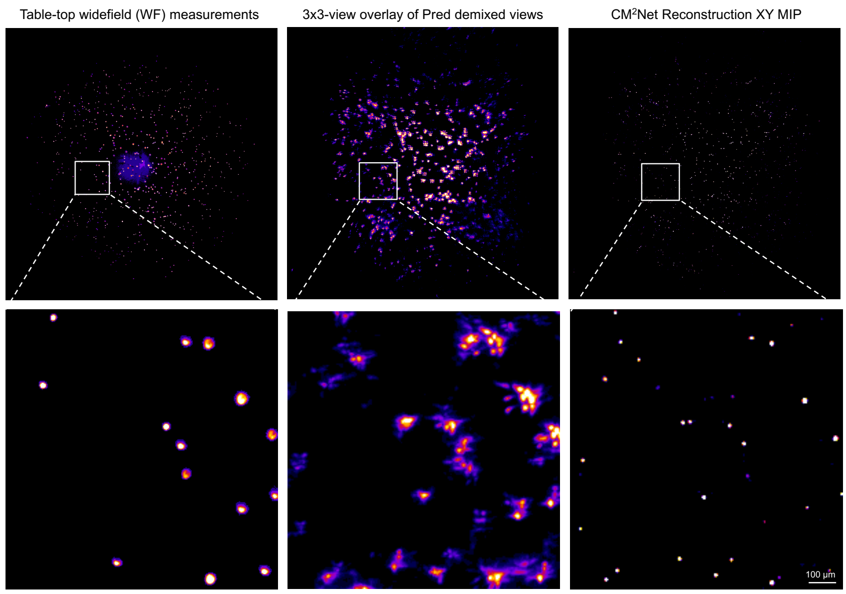

Fig. S18. Comparison of the experimental widefield measurement, 3×3 overlay of predicted view-demixing result, and CM^{2}Net reconstruction MIP. The false positives in the CM^{2}Net reconstruction MIP can find correspondence in the 3×3 overlay predicted demixed views, indicating that the performance decrease for the experimental data originates from the view-demixing net.

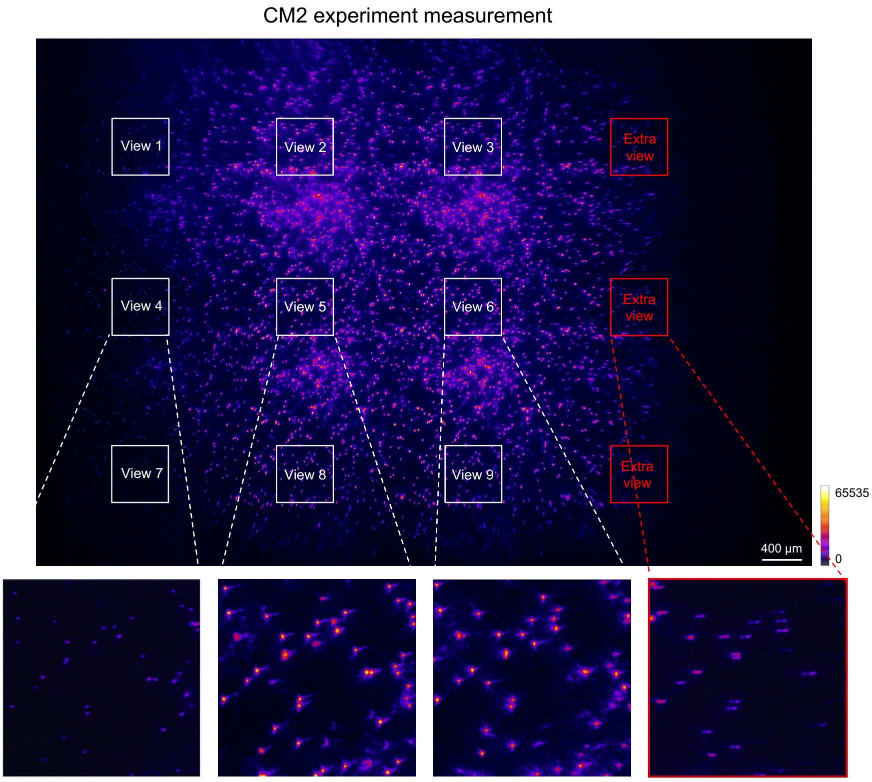

Fig. S19. Extra views in the 10-μm bead experimental measurement. The partial microlenses adjacent to the 3 × 3 microlens array generate extra views in the experimental measurement. The extra views contaminate the 9 central views, which results in increased false positives and mis-detections in the experiments.

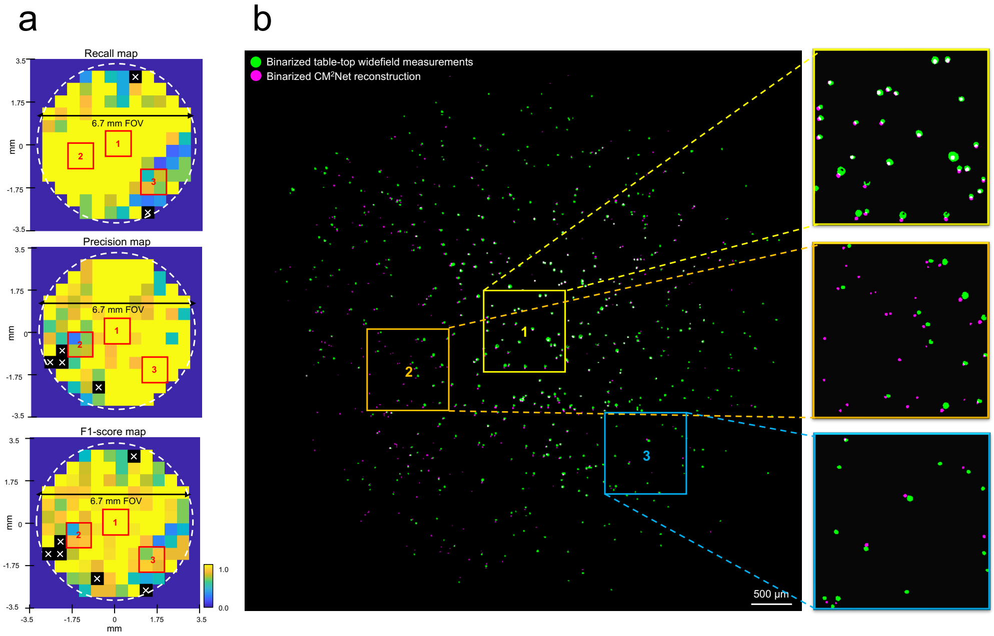

Fig. S20. Quantitative evaluation on 10 µm bead experimental data.(a) The recall, precision, and F1-score maps for the measurement in Fig. 9 of the main text (sample: 10-µm fluorescent beads in a cylindrical volume with 6.7-mm diameter and 0.5-mm depth). The 'x' label indicates the metric value is zero. The dashed circles show the expected boundary of the phantom object. (b) The overlay between the registered and binarized widefield measurement and the CM^{2}Net reconstruction. Patch 1 (size: 1-mm^2) is from the central FOV that achieves the highest F1 = 1.0. In region 1, the binarized map for widefield measurement match perfectly with the CM^{2}Net reconstruction. Patch 2 is from the edge FOV that has a low precision but with perfect recall (i.e. all ground-truth particles are matched but with unmatched falsely reconstructed particles). Patch 3 is from the edge FOV and has a low recall but with perfect precision (all detected particles are matched but with unmatched ground-truth particles). Note that since the image wrapping is unavoidable in the real experimental data and the registration is based only on the central particle pairs, the particle pairs at the peripheral FOV are expected to exhibit slight separations.

For the mixed-size beads phantom in Fig. 8, the recall and precision maps are shown in Fig. S22a. The CM^{2}Net achieves averaged recall ∼0.73 and precision ∼0.84 across the entire 6.5-mm FOV. As compared to the mono-10µm bead experiment, we hypothesize that the decreased recall is attributed to the greater intensity and SNR variations caused by bead size variation in the measurement. The increased precision is due to the reduced FOV and less contaminations from the extra views (see Fig. S21).

For better visualization, we further show the overlays of the registered and paired binarized maps between the widefield measurement and the CM^{2}Net reconstruction. We zoom-in on three patches (size: 1 mm^2) in the metric maps and the corresponding overlay map. Patch 1 from the central FOV achieves the highest F1 score 1.0. The patch with index 2 and 3 is extracted from the edge FOV that either has a low precision with perfect recall (all ground truth particles are matched but with redundant detection) or a low recall with perfect precision (all detected particles are matched but failed to detect some ground truth particles). Patch 2 is from the edge FOV that has a low precision but with perfect recall (i.e. all ground-truth particles are matched but with unmatched falsely reconstructed particles). Patch 3 is from the edge FOV and has a low recall but with perfect precision (all detected particles are matched but with unmatched ground-truth particles). Note that since the image wrapping is unavoidable in the real experimental data and the registration is based only on the central particle pairs, the particle pairs at the peripheral FOV are expected to exhibit slight separations. For direct visualization, we enlarge the corresponding overlap maps in Figs. S20b and S22b. We show in patch 1, the particles detected from the widefield measurement match perfectly with the particles reconstructed by the CM^{2}Net. In patches 2 and 3, we show unpaired particle detected from either the CM^{2}Net reconstruction (low precision) or the widefield measurement (low recall), which are consistent with our metric map, which validates the accuracy of our evaluation procedure.

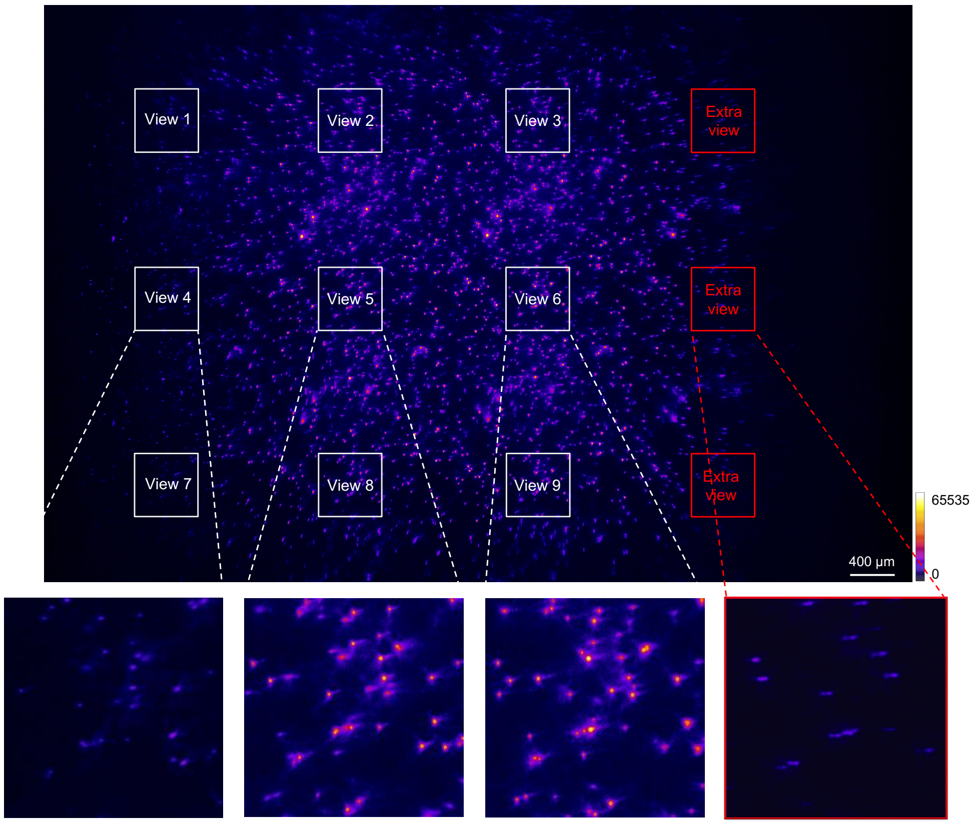

Fig. S21. Extra views in the mixed-size bead experimental measurement. Compared to the mono-10µm bead experiment, the reduced FOV (6.7mm to 6.5mm) reduces the level of contamination from the extra views, which results in decreased false positives and higher precision in the reconstruction.

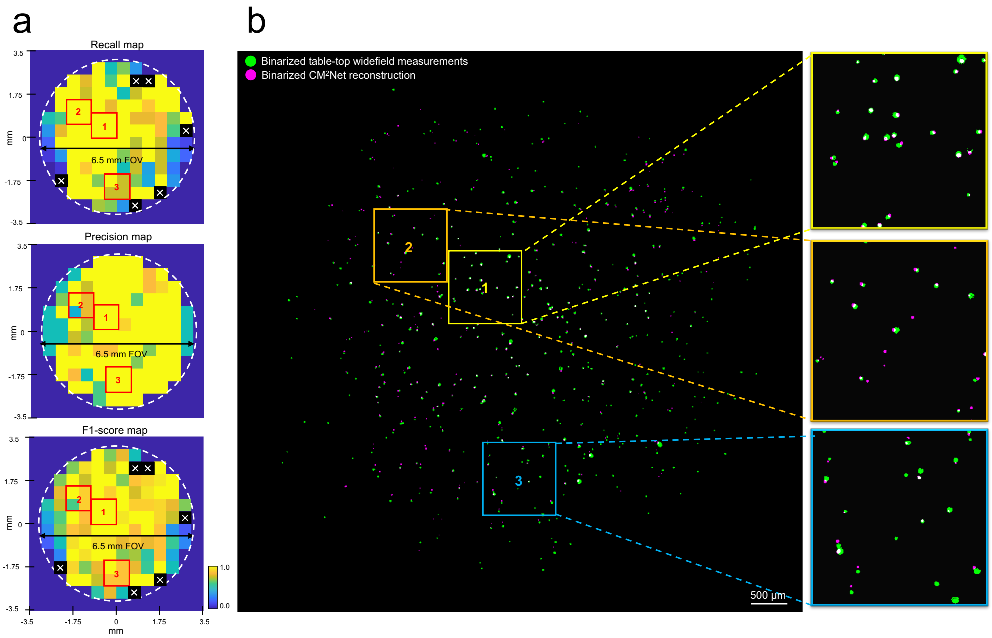

Fig. S22. Quantitative evaluation on mixed-size bead experimental data.(a) The recall, precision, and F1-score maps for the object in Fig. 10 of the main text (sample: 10-μm and 15-μm mixed fluorescent beads in a cylindrical volume with 6.5-mm diameter and 0.8-mm depth). The x labels indicate the metric value is zero and the dashed circles show the expect 6.5-mm diameter region of the phantom object. (b) The overlay of the registered and binarized widefield measurement and CM^{2}Net reconstruction. Patch 1, 2, and 3 (size: 1-mm^2) show example regions of having high F1-score, low recall, and low precision, respectively.

5. NUMERICAL STUDY ON COMPLEX OBJECTS

A. Simulation on cortex-wide 3D neuronal imaging

In this section, we perform pilot numerical study on imaging fluorescently labeled neurons across the entire mouse cortex to show the potential capability of applying CM^2 V2 and the trained CM^{2}Net for large-scale neural imaging. One challenge for in-vivo neural imaging is that the neurons are distributed in the highly curved cortex, which requires the imaging device to be robust to the complex 3D surface geometry. In the main text, we performed thorough analysis on the cylindrical volume embedded with spherical fluorescent emitters. Here, to demonstrate the computational pipeline can be applied to image neurons with irregular shapes, we test CM^{2}Net (trained using the dataset in the main text) and directly apply to simulated CM^2 measurements on the open-source dataset on VIP neurons labeled mouse brain that were imaged by the mesoscale selective plane-illumination microscope (mesoSPIM) [3].

To generate the testing data, we first segment the cerebral cortical region (7.5 × 6.6 × 0.8 mm^3) within a 800 µm depth range from the whole mouse brain volume. We then scale the segmented volume to match with the CM^2 sampling grid. Next, we apply a threshold to remove the fluorescence background and extract the locations and sizes of the connected 3D components (neurons) from the brain volume in MATLAB (`bwconncomp`, `regionprops3`). The original dataset aggregates the neuronal activity across 7-8 min in a single 3D volume, resulting a high neuron density of ∼530/mm^2. We randomly select the neurons from the brain volume to down-sample the number of neurons in each measurement, which simulates the sparse neuronal activity at each time point expected in practice. We simulate two imaging volumes with two neuron densities, including 20/mm^2 and 100/mm^2. Next, we apply the 3D-LSV model on these two volumes to synthesize the CM^2 measurements (as shown in the inset in Figs. S23b and S24b, respectively). Finally, the CM^2 measurement is input to the CM^{2}Net to perform 3D reconstruction.

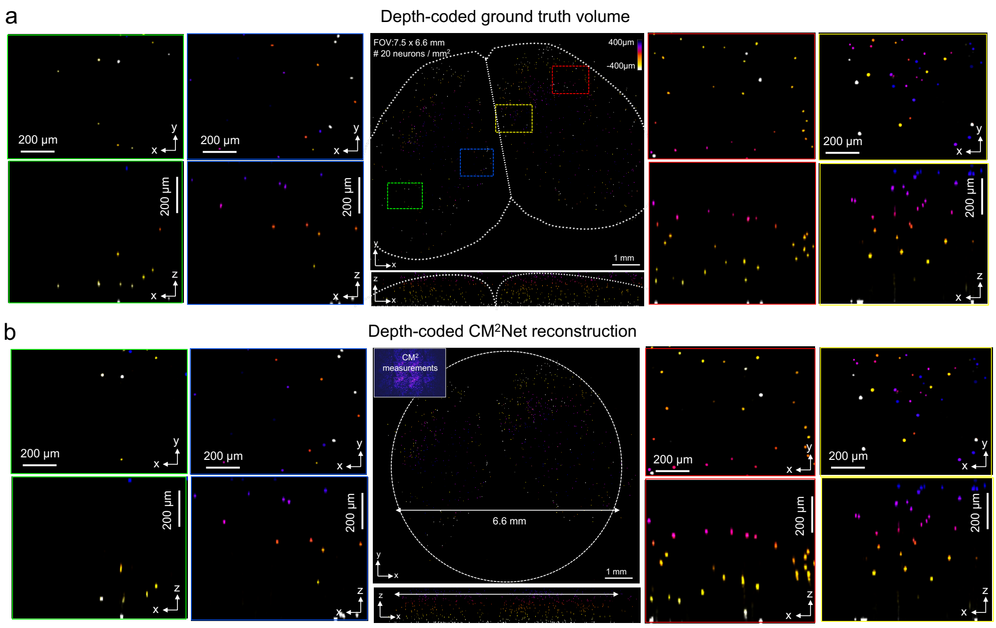

Fig. S23. Numerical study on sparsely labeled neurons in mouse cortex. (a) Visualization of the depth-encoded MIP of the ground truth volume. The volume size is 7.5 × 6.6 × 0.8 mm^2 and the neuron density is around 20 neurons / mm^2. The curved irregular contour indicates the left and right cerebral cortex region. The four patches cross the central and edge FOVs are enlarged for comparison. (b) Visualization of depth-encoded MIP CM^{2}Net reconstruction. The synthesized CM^2 measurement is shown as the inset. At this low neuron density, the CM^{2}Net full-FOV reconstruction is in good agreement with the ground truth.

The imaging volume and CM^{2}Net reconstruction for sparsely labeled neurons is shown in Fig. S23. First, to visualize the curved brain geometry, we register the ground truth volume with the cerebral cortex contour extracted from Allen Mouse Brain Atlas [4](shown by the white dotted line). Next, we assess the full-FOV reconstruction by comparing the depth-encoded MIP of the ground truth volume and the CM^{2}Net 3D reconstruction. We further compare four zoom-in regions from the center (blue and yellow patches) and edge (green and red patches) FOVs. By visual inspection, CM^{2}Net successfully reconstruct the irregular 3D brain geometry and robustly reconstruct the neurons with different intensities and sizes with isotropic high resolution. The reconstruction FOV is limited to around 6.6 mm (labeled by the white circle). There are several miss-detections when exceeds the effective FOV (shown in green patch). These results agree well with the FOV analysis in Section 3.2.

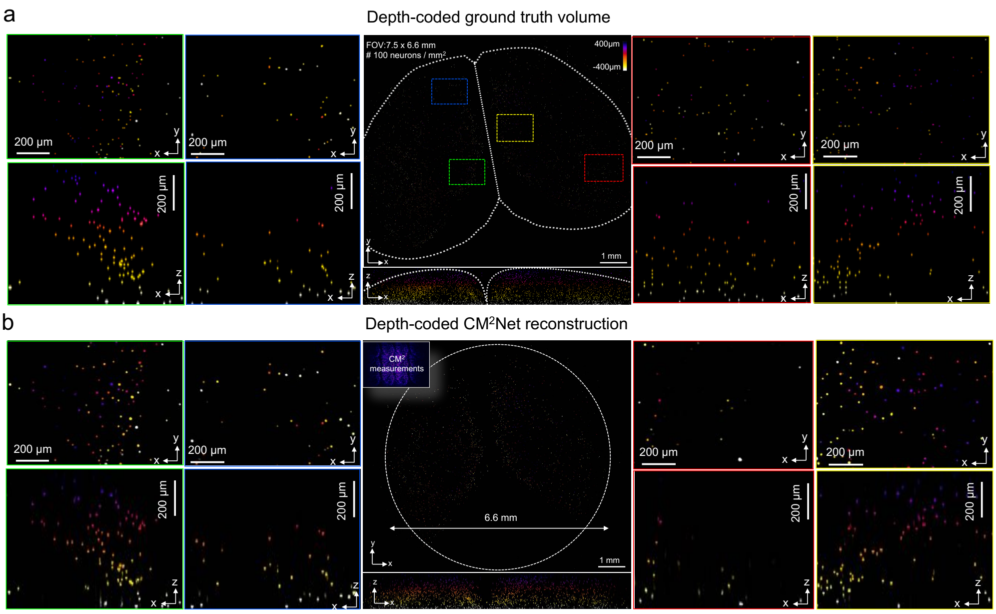

Fig. S24. Numerical study on densely labeled neurons in mouse cortex.(a) Visualization of the ground truth of the imaging volume as the color-encoded MIP. The volume size is $7.5 \times 6.6 \times 0.8$ mm^3 and the neuron density is around 100 neurons / mm^2. The curved irregular contour labels the left and right cerebral cortex region. The four patches at the FOV central and edges are enlarged for comparison. (b) Visualization of the CM^{2}Net reconstruction with depth encoded MIP. The synthesized CM^2 measurement is shown as the inset. At this high neuron density, the CM^{2}Net provides high-quality reconstruction in the central FOV (around 6.6 mm, shown in blue, yellow and green patches). The performance largely degrades at the edges (shown in the red patch), which follows the observation in Section 3.2.

The imaging volume and the CM^{2}Net reconstruction for a densely labeled neuronal volume is shown in Fig. S24. The left and right cerebral cortex contour is labeled in the ground truth volume (white dotted line). First, we validate the full-FOV reconstruction by comparing the XY and XZ MIPs of the depth-coded ground truth volume and the CM^{2}Net 3D reconstruction. To further assess the reconstruction at greater resolution, four regions from the center (green and yellow patches) and corner (blue and red) FOVs are enlarged for comparison. By visual inspection, the reconstruction matches well to the ground truth at the central 6.6×6.6 mm^2 FOV (labeled with white circle). The performance largely degrades when exceeds the effective FOV (example shown in red patch) and it is more severe compared to the low density case, which follows the quantitative analysis in Section 3.2.

B. Simulation on 3D imaging of mouse brain vessels

In this section, we perform a numerical study on a mouse vasculature to demonstrate the ability of CM^2 to image more complex structures. A major concern for deep learning based reconstruction is the lack of generalization ability. Although we trained the CM^{2}Net on only sparse particles to mimic the neuron-like structures, here we study how the CM^{2}Net can be applied to more complex and non-particle objects. We test the CM^{2}Net on the open-source whole mouse brain vessels

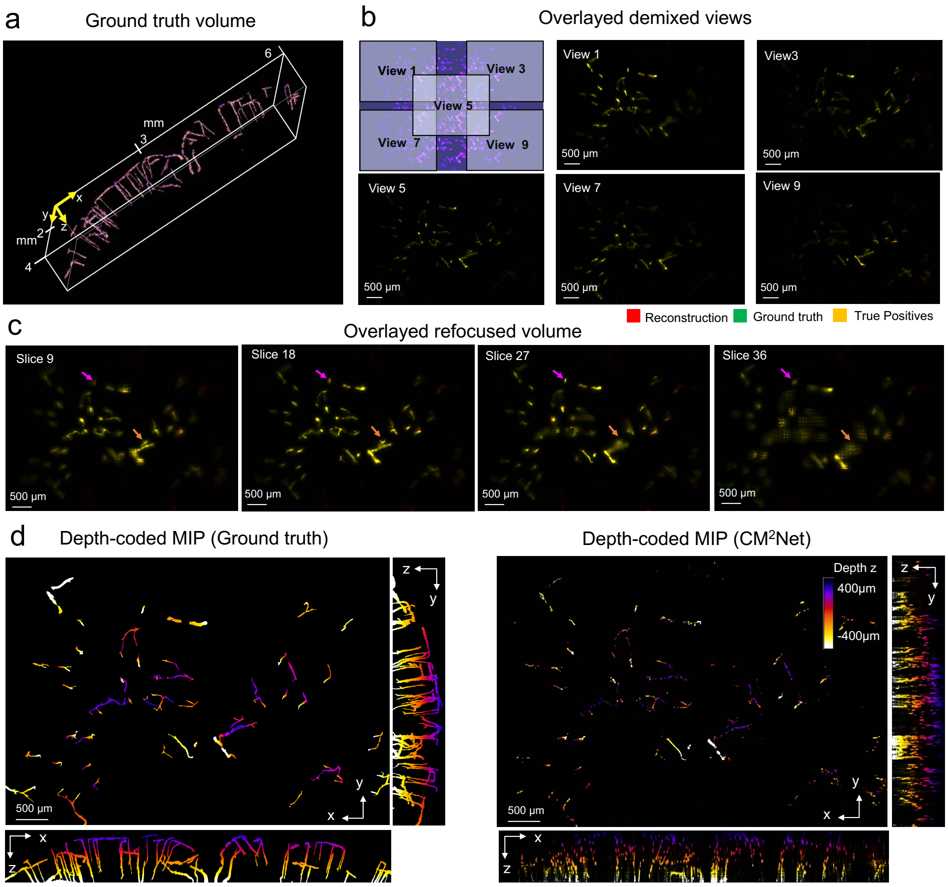

Fig. S25. Numerical study on a mouse vessel network. (a) Visualization of the ground truth volume. The volume size is $6 \times 4 \times 0.8$ mm^3. The small vessels are axially overlapped and distributed in the curved mouse brain. (b) Overlap between the ground truth (green) and network prediction (red) of the demixed views. The first figure is the synthesized CM^2 measurement. The labeled patches are the corresponding input for extracting the demixed views. The result shows that the network view-demixing result is matched well with the ground truth (yellow). (c) Overlap between the ground truth (green) and network prediction (red) of lightfield refocused volume (generated by the shift-and-add operation from the network demixed view stack). The slices shown in the figure are uniformly distributed across the entire refocused depth range. The orange and magenta arrows indicated the small vessels focused at the 18th and 27th slices. The result shows that the refocused volume provides accurate estimation of the 3D information. Results from (b) and (c) show that the particle-trained view-demixing network can generalize well to the continuous dense object. (d) Depth-encoded MIPs of the ground-truth volume and CM^{2}Net reconstruction. The CM^{2}Net can accurately recover the volumetric distribution of the vessels but with discontinuity artifacts due to the sparsity enforced by the particle-trained CM^{2}Net.

segmentation data set [5] to show the particle trained CM^2Net can still recover the volumetric mouse vasculature.

The preprocessing steps are similar to the previous section. First, We segment the cerebral cortical region within the 800 μm depth (volume size: $6\times 4.2 \times 0.8$ mm^3) and extract the vessels from the brain volume. Next, we keep only small vessels within the volume with size ranging from 800-4000 pixels to form the imaging volume. Next, we scale the volume to match the CM^2 sampling grid and simulate the CM^2 measurement by our 3D-LSV model (first figure in Fig. S25b). Finally, the measurement is input to CM^2Net for 3D reconstruction.

The ground-truth volume and the CM^2Net view-demixing and reconstruction results are shown in Fig. S25. First, we show the complexity of the imaging volume in Fig. S25a, where axial overlapped small vessels are distributed in the curved mouse brain. Next, we assess the view-demixing results by visualizing the overlay between the ground-truth demixed views (green) and the network demixed results (red). Figure S25b shows that the network demixed views are matched well with the ground truth (overlay shown in yellow), demonstrating that the view-demixing network is robust to demultiplex the overlapped views for dense continuous objects, even though it is trained entirely on the sparse discrete particle data set. The successful generalization of the CM^2Net view-demixing module lays the foundation for robust reconstruction of the dense object. On the one hand, the demixed views are the images captured from different microlenses, which captures the information within different FOV regions. With an accurately demixed views stack, the VS branch can learn high quality information across the full FOV. On the other hand, the lightfield refocused volume generated from an accurate view stacks can provide focus information for the LFR enhancement branch, which helps the enhancement branch to accurately recover the object's depth information. To demonstrate this, we show two small vessels digitally refocused at different depth in Fig. S25c. Figure S25c shows the highly matched overlay between the ground truth and the predicted lightfield refocusing volumes at different depth planes. Finally, we validate CM^2Net final Reconstruction by comparing the XY, XZ and YZ MIPs of the depth-encoded ground-truth and CM^2Net reconstruction in Fig. S25d. By visual inspection, the CM^2Net reconstruction matches well with the ground truth in all projections, showing that the CM^2Net can preform high-quality volumetric reconstruction for such dense object. However, we also observe discontinuity artifacts in the reconstruction result. This is attributed by the fact that the CM^2Net was only trained on the sparse particle objects, which implicitly enforces a sparsity prior on the reconstruction volume. We expect one can remove the discontinuity artifacts by fine tuning the CM^2Net on a more complex training dataset.

REFERENCES

1. Y. Xue, I. G. Davison, D. A. Boas, and L. Tian, "Single-shot 3d wide-field fluorescence imaging with a computational miniature mesoscope," Sci. advances **6**, eabb7508 (2020).
2. Y. Bao, S. Soltanian-Zadeh, S. Farsiu, and Y. Gong, "Segmentation of neurons from fluorescence calcium recordings beyond real time," Nat. Mach. Intell. **3**, 590–600 (2021).
3. F. F. Voigt, D. Kirschenbaum, E. Platonova, R. A. Campbell, R. Kastli, M. Schaettin, L. Egolf, A. Van der Bourg, P. Bethge, K. Haenraets *et al.*, "The mesospim initiative: open-source light-sheet microscopes for imaging cleared tissue," Nat. methods **16**, 1105–1108 (2019).
4. E. S. Lein, M. J. Hawrylycz, N. Ao, M. Ayres, A. Bensinger, A. Bernard, A. F. Boe, M. S. Boguski, K. S. Brockway, E. J. Byrnes *et al.*, "Genome-wide atlas of gene expression in the adult mouse brain," Nature **445**, 168–176 (2007).
5. M. I. Todorov, J. C. Paetzold, O. Schoppe, G. Tetteh, S. Shit, V. Efremov, K. Todorov-Völgyi, M. Düring, M. Dichgans, M. Piraud, B. Menzeo, and A. Ertürk, "Machine learning analysis of whole mouse brain vasculature," Nat. methods **17**, 442–449 (2020).