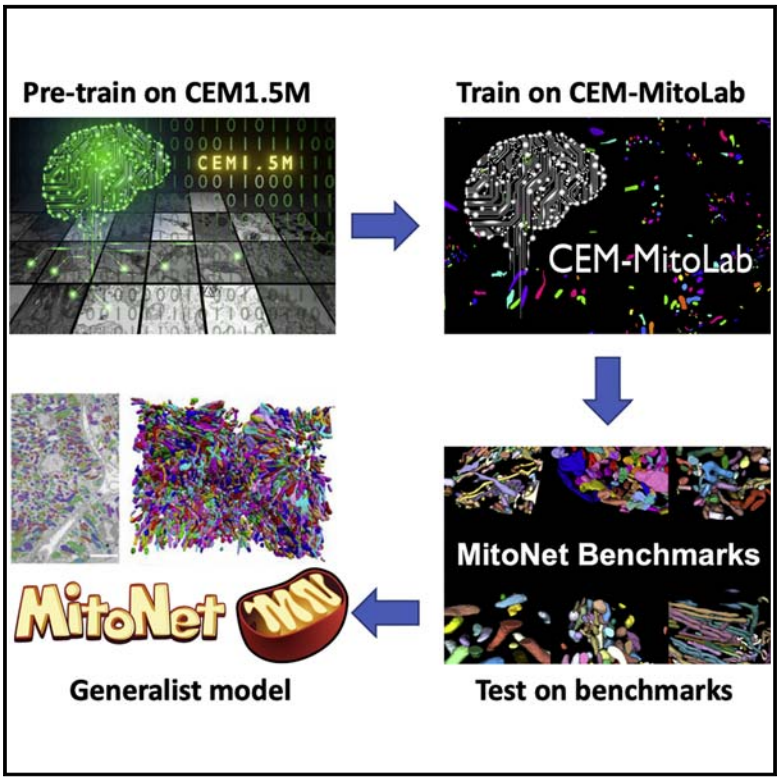
Cell Systems

Instance segmentation of mitochondria in electron microscopy images with a generalist deep learning model trained on a diverse dataset

Graphical abstract



Authors

Ryan Conrad, Kedar Narayan

Correspondence

kedar.narayan@nih.gov

In brief

To rapidly segment mitochondria from any given electron microscopy (EM) image, Conrad and Narayan curate massive EM datasets to train a generalist deep learning model, MitoNet. Through an accessible GUI, empanada-napari, MitoNet generates accurate segmentation results on various 2D and 3D EM images, enabling easy visualization and quantitation.

Highlights

- A curated dataset of ~1.5 million diverse, unlabeled electron microscopy images
- ~135K segmented mitochondria from existing datasets and crowdsourced annotations
- MitoNet, a deep learning model, accurately segments mitochondria in various contexts
- Empanada, a Python library and napari plugin, allows MitoNet inference and cleanup

Conrad & Narayan, 2023, Cell Systems *14*, 58–71
January 18, 2023
https://doi.org/10.1016/j.cels.2022.12.006

CellPress

CellPress

Cell Systems

Methods

Instance segmentation of mitochondria in electron microscopy images with a generalist deep learning model trained on a diverse dataset

Ryan Conrad[1,2] and Kedar Narayan[1,2,3,*]

[1]Center for Molecular Microscopy, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892, USA

[2]Cancer Research Technology Program, Frederick National Laboratory for Cancer Research, Frederick, MD 21702, USA

[3]Lead contact

*Correspondence: kedar.narayan@nih.gov

https://doi.org/10.1016/j.cels.2022.12.006

SUMMARY

Mitochondria are extremely pleomorphic organelles. Automatically annotating each one accurately and precisely in any 2D or volume electron microscopy (EM) image is an unsolved computational challenge. Current deep learning-based approaches train models on images that provide limited cellular contexts, precluding generality. To address this, we amassed a highly heterogeneous $\sim 1.5 \times 10^6$ image 2D unlabeled cellular EM dataset and segmented ~135,000 mitochondrial instances therein. MitoNet, a model trained on these resources, performs well on challenging benchmarks and on previously unseen volume EM datasets containing tens of thousands of mitochondria. We release a Python package and napari plugin, empanada, to rapidly run inference, visualize, and proofread instance segmentations. A record of this paper's transparent peer review process is included in the supplemental information.

INTRODUCTION

Electron microscopy (EM) reveals high-resolution snapshots of cellular and subcellular ultrastructure. Recent volume EM advances have enabled 3D imaging[1,2] of increasingly large specimens, notably in connectomics[3–5] where deep learning (DL) algorithms are actively employed to generate wiring diagrams and enable quantitative analyses.[6–9] Similar methods have also been used to investigate organellar structures in other systems,[10–14] generating biological insights at unprecedented scales. The typical DL workflow for these applications is to densely annotate features in a 3D region of interest (ROI), train a model on these annotations, run inference on the full dataset, and then proofread the model output.[6,7,10,11,15,16] Although visually impressive, these results belie a failure to generalize, meaning that segmentation quality drops dramatically when models are presented with images from unseen cellular milieus or different sample preparation or imaging protocols.[10,11,17] Poor model generalization thus forces repeated cycles of annotate, train, infer, and proofread for every new project—this laborious workflow could be drastically simplified and accelerated by the use of a generalist DL segmentation model.

As ubiquitous and morphologically complex organelles that play critical roles in cellular physiology and pathology,[18–26] mitochondria provide both a stringent test and a high payoff for such a model. Studies of key aspects of mitochondrial biology such as morphologies,[27–30] networks,[31,32] and fission-fusion cycles[33,34] would benefit from the high-resolution 3D imaging afforded by volume EM. But robust quantitation from these experiments requires precise and accurate labeling of each of the hundreds of "instances" per cell, and the solution must be efficient and general to process large and varied volume EM datasets. Fortunately, despite their extraordinary variety, mitochondria are instantly recognizable by their ultrastructure, hinting at the potential for a "universal" DL model that can accurately and precisely recognize them in any EM image. However, their heterogeneity and the vast cellular landscape within which they are presented means that (1) this task is fundamentally different from neuronal tracing and (2) the model may not be well trained by homogeneous datasets such as MitoEM,[15] currently the largest available labeled dataset, derived from brain tissue only. We hypothesized that sparse 2D instance segmentations from an eclectic set of EM images could effectively expand the range of contexts needed for model generalization, and at a much lower cost than dense 3D segmentation.

Here, we release the following resources: we have curated cellular EM (*CEM*)*1.5M*, a heterogeneous, non-redundant, information-rich, and relevant unlabeled EM image dataset (at $\sim 1.5 \times 10^6$ images, the largest of its kind to the best of our knowledge) for use as a database to pre-train and sample images for mitochondrial, or other organelles, segmentation. Combining existing labeled datasets and crowdsourced annotations of images from CEM1.5M, we have created *CEM-MitoLab*, a similarly diverse dataset for training mitochondrial segmentation models.

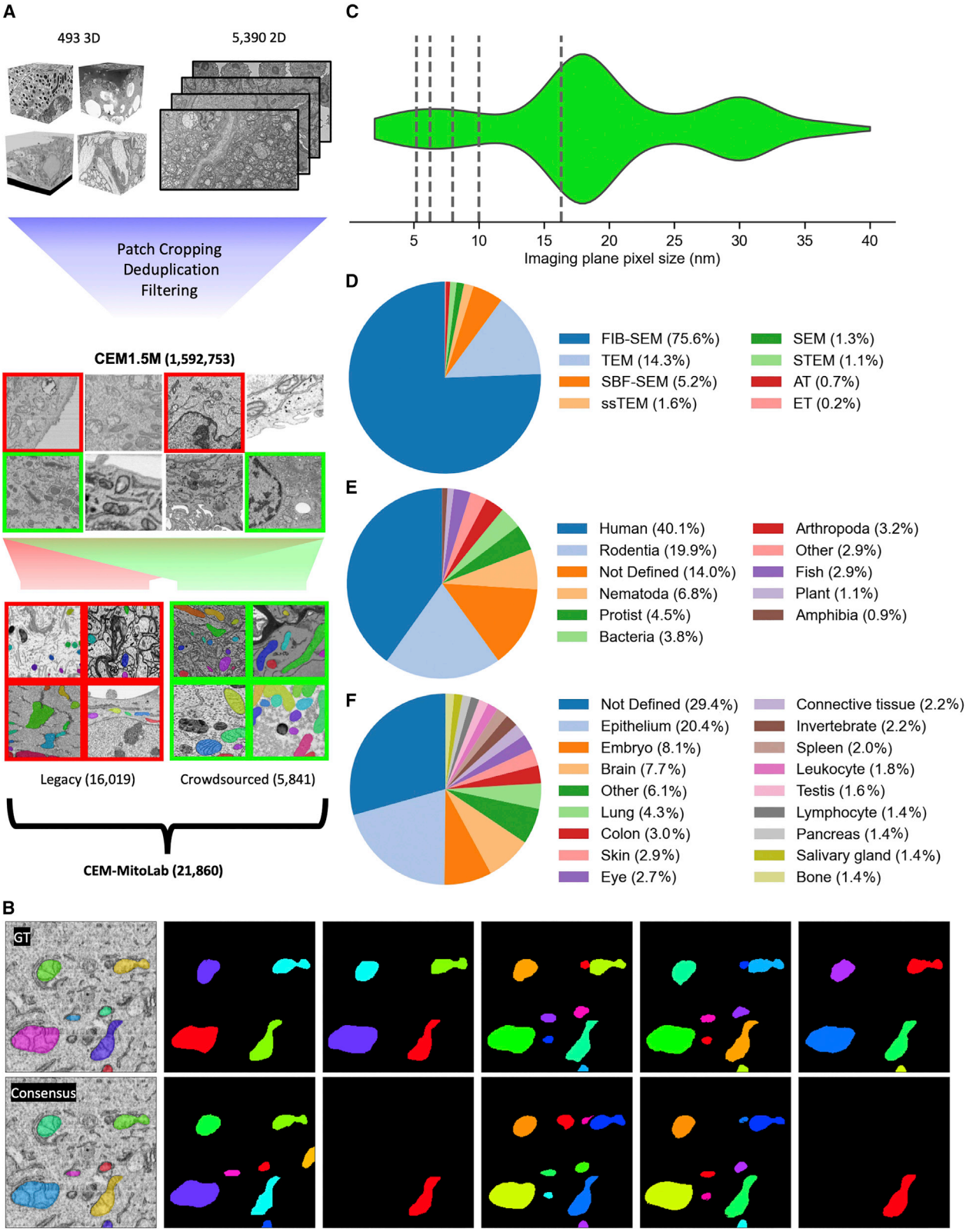


Figure 1. Creation of a diverse and representative dataset for mitochondrial instance segmentation

(A) Schematic of the data curation pipeline. Volume EM reconstructions and 2D EM images were curated to create CEM1.5M. Random patches from previously labeled data (legacy annotations, red) and crowdsource annotated patches from CEM1.5M (green) were combined to form CEM-MitoLab.

(legend continued on next page)

We show that *MitoNet*, a Panoptic-DeepLab model trained on CEM1.5M and CEM-MitoLab, outperforms equivalent models trained on other candidate datasets when tested on challenging 2D and 3D "*MitoNet Benchmarks*." Finally, we use *empanada*, a new Python package and napari plugin for model training, inference, fine-tuning, and proofreading to efficiently create tens of thousands of high-quality mitochondrial segmentations from public volume EM datasets of mouse liver and kidney tissue and make visual and quantitative comparisons between them.

RESULTS

CEM1.5M and CEM-MitoLab dataset creation

Our overall curation process is outlined in Figure 1A. We first generated a database of unlabeled EM images of cells and tissues comprising 466 volume EM datasets (356 in-house, 110 externally generated and publicly deposited, total > 2 PB). To limit overrepresentation, datasets larger than 5 GB were cropped into random 3D ROIs while smaller datasets were retained as-is. This yielded 15,152 3D ROIs equaling 338 GB. To these, we added 28 videos of EM stacks from online publications and 12,658 traditional 2D transmission EM (TEM) and scanning transmission EM (STEM) images (5,657 in-house and National Cancer Institute (NCI) EM data; 7,001 external). Metadata and attribution were recorded, where possible, for all datasets (Data S1). We then applied our previously developed data preparation and curation pipeline[35] to sample xy, xz, and yz planes of isotropic volumes, remove near-duplicate patches within datasets, and triage "uninformative" patches as classified by a neural network. This process created 1,592,753 unique 2D image patches, called *CEM1.5M*.

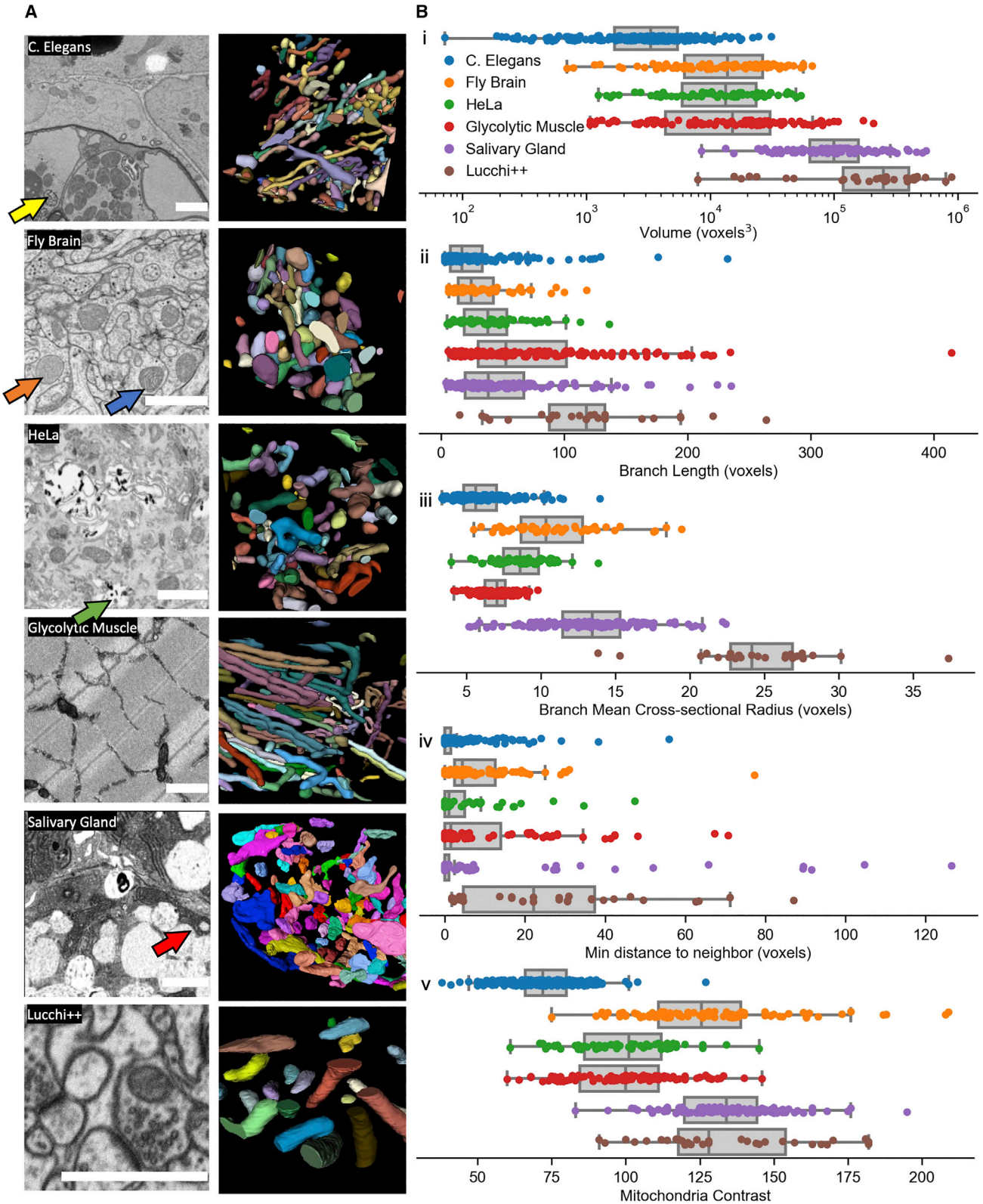
Of the combined 494 3D and video datasets, we selected 489 for annotation, excluding volumes with severe artifacts or >40 nm pixel size. From 19 of the volumes with mitochondrial instances already segmented (7 in-house, 12 external), 12,622 cropped 2D patches were combined with 3,397 other 2D in-house segmentations from 21 datasets to form the "legacy" dataset. Inspired by the citizen science project Etch-a-cell,[13] we used the Zooniverse platform (Figure S1A) to crowdsource the annotation of 5,481 additional images to 34 briefly trained local high school students. To evenly represent the collected data in the subset used for crowdsourcing (Figure S2A), we sampled a maximum of 15 random patches from every 3D dataset and up to 25 patches from each of 83 directories containing 2D datasets. Each image was annotated independently by ten students (Figures 1B and S1B), and we developed a robust algorithm to combine these annotations into a consensus instance segmentation (Figure S3). All consensus segmentations were reviewed by at least two experienced researchers to create the final ground truth. The total 21,860 annotated images constitute *CEM-MitoLab* and contain 135,285 mitochondrial instances. They represent myriad image sizes, pixel resolutions, imaging techniques, sample preparation protocols, cells, tissues, and organisms (Figures 1C–1F, S2B, and S2C). Our dataset consists exclusively of 2D images and favors a broad but superficial sampling of numerous cellular contexts in comparison to the more homogeneous appearance of mitochondria in singly sourced datasets[10,15] (Figures S2D and S2E).

Benchmark dataset characterization

To assess how well models trained on CEM1.5M and CEM-MitoLab generalize to unseen images, we withheld six volumes and 100 2D TEM images from the above pipelines and created mitochondrial segmentations for each (Figures 2A and S4A). The volumes included conventionally fixed, heavy metal stained, and resin-embedded samples of the fly brain,[36] mouse brain,[37,38] glycolytic muscle,[32] mouse salivary gland, and HeLa cell in addition to a high-pressure frozen (HPF), freeze substituted *C. elegans*, all independently imaged by focused ion beam scanning EM (FIB-SEM) and with isotropic-voxel sampling. The first three of these were externally generated, and the last three were in-house. The mouse brain dataset was the well-studied Lucchi++ benchmark test set with the original semantic segmentations converted to instance segmentations. (Note, the Lucchi++ training set was excluded from CEM1.5M and CEM-MitoLab.) Lucchi++ has 5 nm voxels but we intentionally chose or resampled the other volumes to have 10–25 nm resolutions (fly brain 12 nm, HeLa 15 nm, salivary gland 15 nm, glycolytic muscle 18 nm, and *C. elegans* 24 nm)—sufficient to easily identify mitochondria, and in line with the most common resolutions in the training dataset. The TEM image benchmark, comprising 65 images from internal EM data and 35 from external sources, served as a stringent test of 2D model generalization. Although metadata was not available for all images, at least six kinds of organisms and 10 kinds of tissue were sampled and the mitochondrial instances presented a range of contrasts, shapes, sizes, and spatial distributions (Figure S4).

Similarly, the benchmark volumes encapsulate different mitochondrial morphologies and varying levels of difficulty. We skeletonized and created meshes for all mitochondria to make quantitative comparisons of volume, length, and cross-sectional radius along branches in the skeleton, the minimum distance between a mitochondrial mesh and its nearest neighbor, and the difference between the brightest and darkest voxel in each mitochondrion (Figures 2Bi–2Bv), The simple mitochondria in the fly brain volume have distinct variants: lightly stained with poorly defined cristae and darkly stained with well-defined cristae (Figure 2A, orange and blue arrows, respectively). The HeLa cell volume is cluttered with organelles and vesicles and has localized heavy metal precipitates (green arrow). The *C. elegans* volume has an overall lower contrast (Figure 2Bv), as expected for an HPF sample, and presents difficult-to-segment mitochondrial morphologies—small puncta and skinny tubules (Figures 2Bi–2Biii). There are also membranous organelles[39] that appear similar to mitochondria with swollen cristae at these resolutions (yellow arrow). The glycolytic muscle volume has a relatively uncluttered background but complex and elaborately branched morphologies. The salivary gland volume is the most challenging because of flat and bowl-shaped mitochondria tightly pressed

(B) Example of crowdsourced annotation with ground truth (GT, top left), consensus annotation (bottom left), and ten independent student annotations of a representative image showing a high degree of consensus.
(C–F) Dataset distribution by various parameters in CEM-MitoLab. (C) Imaging plane pixel sizes of volume EM images (N = 489). Dashed lines, 2D EM images. (D) Imaging technique, (E) source organism, (F) source tissue (vertebrates only, *in vitro* cells grouped under not defined; N = 593).



(legend on next page)

against salivary granules (red arrow), weak staining, and close packing of the mitochondria themselves (Figure 2Biv).

Deep learning model and post processing

For 2D instance segmentation, we adopted Panoptic-DeepLab (PDL).[40] PDL is an encoder-decoder architecture, like the standard U-Net,[41] that takes a bottom-up approach to instance segmentation and scales to segment arbitrarily large numbers of objects within a field of view (top-down and direct set prediction algorithms like Mask R-CNN[42] and Mask2Former[43] cap the number of detections). We trained PDL models to infer mitochondrial semantic segmentations, centers, and per-pixel x and y offsets from the associated object's center (Figure 3A). To create segmentations at resolutions higher than the input image resolution, we employed PointRend,[44] which iteratively interpolates and reevaluates the label at the most uncertain pixels. This feature allows the model to accept downsampled images and output precise segmentation boundaries at the original resolution without a resolution-specific model architecture (Figure S5). The semantic segmentation, offsets, and object centers were post processed into an instance segmentation.

To generalize model inference to three dimensions, we tracked 2D instance predictions through a volume EM stack by performing a 1-to-1 matching of objects across consecutive slices using the Hungarian algorithm.[45] We also adopted two post-processing steps to counter oversplitting errors common for this approach. First, we computed intersection-over-area (IoA) scores for objects that were unmatched because of the 1-to-1 assumption. Unmatched objects with IoA scores over a threshold value were merged to the label of the object with which they overlapped in the preceding slice (Figure 3B). Second, we calculated 2D instance segmentations using median semantic segmentation probabilities over neighborhoods of three or more slices to cover gaps and produce smooth and continuous objects in 3D (Figure 3C). Finally, we performed matching in the reverse direction through the stack to account for branched morphologies (Figure 3D). Optionally, for isotropic-voxel volumes, *ortho-plane inference*[46] can be applied. Here, the inference is performed independently on stacks of xy, xz, and yz images and the predictions are combined into a single consensus segmentation using the same algorithm developed for our crowdsourced annotations (Figure 3E). IoA-based merging, median filtering, and ortho-plane inference steps all yielded significant improvements in the volumetric benchmark datasets (Table S1).

The particulars of this model architecture and post-processing scheme were chosen to balance usability and segmentation performance. As we describe later, we can quickly segment large 3D datasets with limited compute resources because only a few 2D image planes are ever held in memory at a time. However, we also experimented with two other methods. The first was a model based on EfficientDet[47] that has half the parameters of PDL but uses the same prediction task and 2D and 3D post processing. The second was a UNet-BC as implemented in PyTorch Connectomics[48] and previously developed for the MitoEM benchmark. Briefly, this model predicts two semantic segmentations: one for the bulk of the mitochondrion and another for the outer boundary. These predictions are post processed into an instance segmentation using the watershed algorithm applied in 3D.

MitoNet training and evaluation

The encoder network in our experiments was a ResNet50[49] pre-trained on CEM1.5M using the SwAV[50] self-supervised learning algorithm (the effect of pre-training is summarized in Table S2). Our best-performing PDL model, *MitoNet*, was trained for 120 epochs on CEM-MitoLab. On the benchmark datasets, we observed that it segmented a variety of mitochondrial instances accurately in 2D and 3D (Figures 4A and 4B). MitoNet performed best on the relatively simple brain tissue datasets (fly brain and Lucchi++) with both semantic IoU and F1@75 scores around 0.9 or higher (Figure 4C). Detailed performance metrics are included in Table 1. MitoNet's F1@75 score of 0.88 on the Lucchi++ test set matched specialist models that were trained exclusively on the Lucchi++ training data.[51] Despite the heterogeneity of the TEM benchmark and that only 16% of CEM-MitoLab was composed of TEM images, the model reached an average F1@75 score of 0.68 (unlike the other benchmarks F1 was calculated in 2D). Although a small subset of images in the TEM benchmark was not well segmented, more than 75% had F1@50 scores greater than 0.6, and the median F1@50 was greater than 0.8 (Figure S6B).

On the more difficult HeLa and glycolytic muscle benchmarks, MitoNet achieved semantic IoU scores of 0.79 and 0.84 but lower F1@75 scores of 0.5 and 0.6. For the HeLa benchmark, the model successfully ignored much of the cluttered intracellular features, but we did observe several false positive (FP)s caused by Golgi (data not shown). The IoU and F1 scores on the glycolytic muscle dataset were high given that only a handful of training images were from muscle tissue and none were from glycolytic muscle (whose mitochondrial ultrastructure is markedly different[32]). Although MitoNet correctly detected 78% of closely apposed mitochondria (i.e., <5 voxels to nearest neighbor) at an IoU threshold of 0.5 in the HeLa volume and 52% in the glycolytic muscle, visual inspection and the lower instance compared with semantic segmentation scores suggest that identifying individual mitochondria in crowds remains a challenge (Figure S6A, min distance plot). Figure 4B shows examples of accurate mitochondrial segmentations, including a heavily branched mitochondrion, as well as overmerging and oversplitting errors that our method was prone to make—such errors could substantially lower F1 scores even when pixel-level accuracy was high (black arrow).

Figure 2. Challenging and diverse volume EM benchmarks for evaluating automatic instance segmentation performance

(A) 2D representative images (left) and 3D reconstructions (right) for the benchmark test sets. Top to bottom: *C. elegans*, fly brain, HeLa cell, glycolytic muscle, salivary gland, and Lucchi++. Yellow arrow, membranous organelle; orange and blue arrows, lightly and darkly stained mitochondria; green arrow, heavy metal precipitate; red arrow, mitochondrion and tightly apposed salivary granule in the acinus.

(B) Comparison of individual mitochondria and boxplots across benchmarks by (top to bottom): (i) volume (log scale), (ii) branch length, (iii) mean cross-section radius, (iv) minimum distance to neighbor (all in voxels), and (v) mitochondrial contrast. Blue, *C. elegans* N = 241; orange, fly brain N = 91; green, HeLa cell N = 68; red, glycolytic muscle N = 104; purple, salivary gland N = 131; brown, Lucchi++ N = 33. Scale bars, 1 μm.

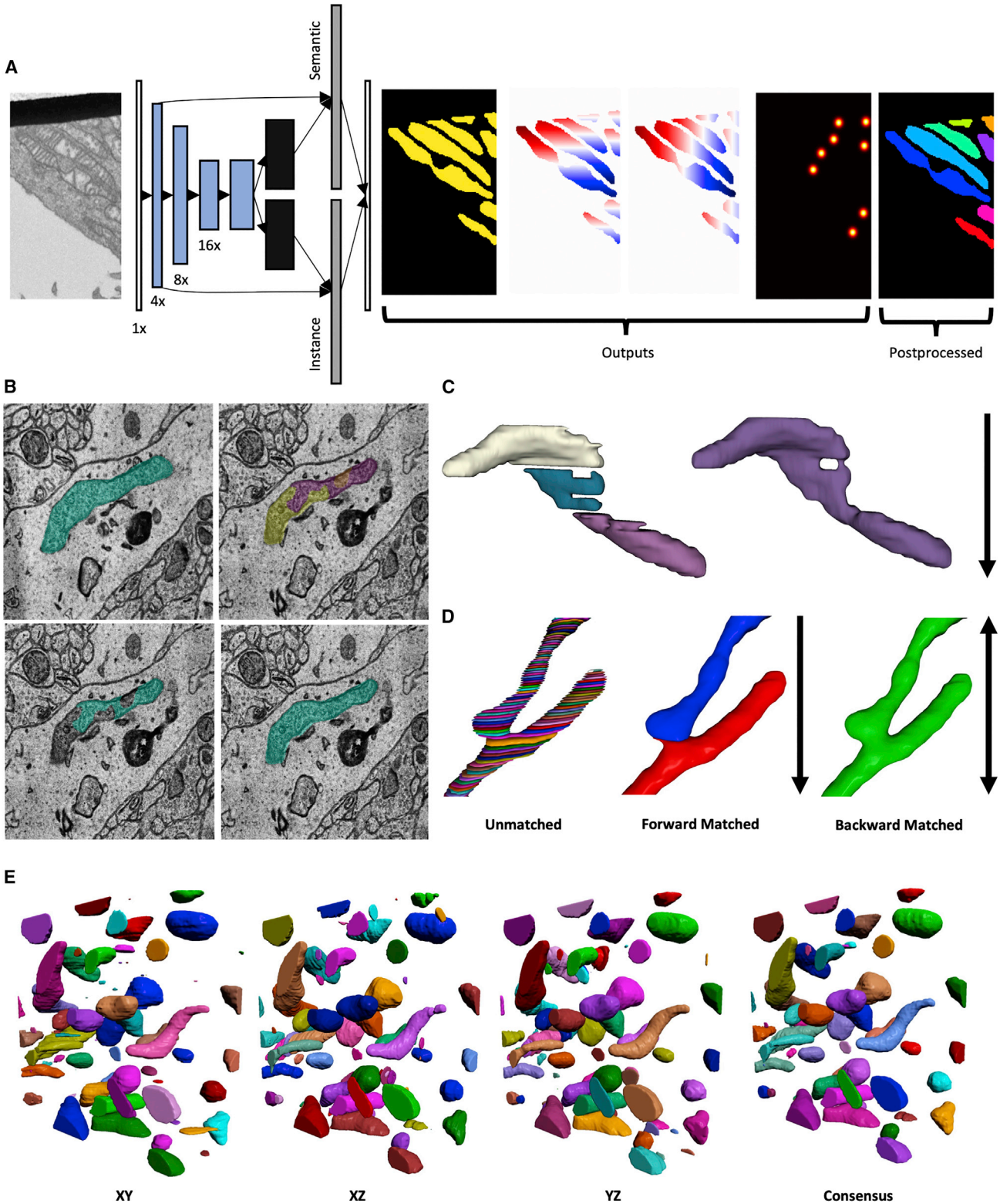


Figure 3. Deep learning model and post-processing pipeline to create 2D or 3D instance segmentations

(A) Schematic of Panoptic-DeepLab showing the input grayscale image (left, blue boxes, encoder layer outputs; black boxes, atrous spatial pyramid pooling (ASPP) layer outputs; gray boxes, decoder layer output). Outputs of the network are (left to right) semantic segmentation, up-down offsets, left-right offsets, and the instance centers heatmap. Far right, instance segmentation created from the outputs.

(legend continued on next page)

The *C. elegans* dataset was challenging because of the small and low-contrast mitochondria that were tightly packed together. Here, we used a lower vote threshold of only one plane for ortho-plane inference to increase IoU from 0.44 to 0.60 and F1@75 from 0.18 to 0.33 (Table S1). MitoNet struggled to detect mitochondria smaller than 1,000 voxels ($\sim$0.25 μm^3)—these accounted for about 25% of all false negative (FN)s (Figure S6). The model also performed slightly worse on mitochondria with lower contrast (mean grayscale of true positive (TP), 75.9; FN, 70.5, $p < 0.001$; see Table S3 for complete statistics). Closely apposed instances accounted for over 75% of all FNs at an IoU threshold of 0.5. The model mostly avoided erroneously labeling the membranous organelles but occasionally labeled some vesicles as mitochondria (Figure 4A, yellow arrow). On the most challenging benchmark, the salivary gland, MitoNet achieved an IoU score of just 0.10 and F1@50 of 0.04.

Our smaller EfficientDet-based model, *MitoNet-mini*, had an average IoU score across the benchmarks that was only 4.5% worse than MitoNet while using 66% less graphics processing unit (GPU) memory per image (Figures S6C; Table S4). There was a sharper decline in the average F1@75 score (13.5%), which suggests that the model did not have enough capacity to fit the harder instance segmentation task. On the other hand, the UNet-BC model outperformed MitoNet on both IoU and F1@75 scores by 4.0% and 10.2%, respectively (Figure S6C; Table S5). However, this came at the cost of nearly 4× more GPU memory per image, and for the largest benchmark dataset of just 1.5 GB, it required 28 GB of memory to post process.

Although MitoNet can reliably provide automatic segmentations in a variety of contexts, it is possible to improve its performance on a particular dataset via fine-tuning. For each volume EM benchmark, we took a small sample (n) of patches constituting <2% of all voxels (n = 5 for *C. elegans*, HeLa, fly brain; n = 8 for glycolytic muscle; n = 16 for Lucchi; n = 64 for salivary gland) and minimally fine-tuned MitoNet for 100 epochs. Fine-tuning improved IoU scores by 19.8%, with all but the salivary benchmark exceeding IoUs of 0.8; F1@75 scores improved by 10.4% (Figure 4D; Table S6). This underscores an important point: even when the generalist MitoNet model fails on a dataset, it is still a strong starting point to train a specialist model with a modest number of examples and compute time. A movie describing the procedure of model fine-tuning with empanada is also included (Video S1)

Finally, we directly probed the efficacy of CEM-MitoLab against other training datasets by measuring the performance of models across all benchmarks except the salivary gland volume. We trained models for approximately 10,000 iterations to account for the different number of images in each dataset (for comparison, the 120 epochs used for MitoNet were equivalent to about 40,000 iterations). The model trained on the Heinrich et al.[10] dataset performed poorly on the benchmarks, likely because of its small size and limited breadth. Training on MitoEM, a large but homogeneous neuronal dataset, gave F1@75 scores of 0.86 and 0.75 on the fly brain and Lucchi++ volumes, but scores of 0.25, 0, 0.07, and 0.01 on the TEM, *C. elegans*, HeLa cell, and glycolytic muscle benchmarks. The model trained on expert-corrected crowdsourced annotations was almost matched by a model trained on our legacy dataset, a 5× larger, reasonably heterogeneous sampling of 19 volumes, over 500 TEM images, and over 2,000 randomly chosen 2D patches from CEM500K. In both cases, IoU and F1 scores were substantially better across all benchmarks than training on MitoEM (Figure 4D; Table 2). Indeed, even training on the *uncorrected* student consensus annotations was better than training on MitoEM for all benchmarks except the fly brain volume, emphasizing how critical data diversity is, even at the expense of per-instance accuracy (Data S2). Sparsely sampled, noisy, but heterogeneous labeled data can be more effective for training generalist segmentation models than much larger densely annotated, expert-proofread, but homogeneous data. Our best results were achieved by training on CEM-MitoLab (i.e., the combination of legacy and expert-proofread crowdsourced annotations).

Analyzing mitochondrial morphologies in mouse kidney and liver tissue

Mitochondria are enriched in and critical to kidney function especially in proximal tubules, and their dysfunction marks renal disease and injury.[24–26,52] Despite the centrality of mitochondria in renal pathology, existing studies are limited by a reliance on light microscopy, 2D EM, or semantic segmentations of limited volume EM data.[53–55] To demonstrate the power of MitoNet for general mitochondrial instance segmentation and generate 3D ultrastructural correlates of known differences in mitochondrial energetics between proximal and distal tubules,[56] we used the same model to automatically segment tens of thousands of mitochondria in two previously unseen volumes from mouse kidney tissue downloaded from OpenOrganelle.[36] The ROIs contained primarily distal tubules and proximal tubules (10 GB each). Additionally, an ROI of hepatocytes from liver tissue (30 GB) was cropped from another volume for comparison. The full python package that we created to run MitoNet inference as well as the post-processing pipeline is called empanada, while a user-friendly napari plugin of the same name allows easy graphical user interface (GUI)-based inference and post-processing, point-and-click merge and split operations, and model fine-tuning (Video S2). MitoNet segmented 7,718, 15,180, and 61,244 objects in each ROI, respectively. On a system with a single GPU and just 16 GB of random access memory (RAM), automatic segmentation of all objects with empanada took 1.5 h for each of the kidney volumes and 3.5 h for the liver volume (watershed post processing, like that used for UNet-BC, required 512 GB of RAM). Out-of-the-box, MitoNet predictions were

(B) Instance matching across adjacent slices uses intersection-over-union (IoU) and intersection-over-area (IoA) scores. Clockwise from the top left: predicted segmentation of slice j, j + 1, IoU, and IoA merging, IoU only merging.
(C) Result of median filtering in the direction of the black arrow.
(D) From left to right: stacked 2D segmentations of before matching, after forward matching only, and after forward and backward matching. Black arrows denote the direction of matching.
(E) An example of 3D instance segmentation of mitochondria after running inference in (left to right) xy, xz, yz directions, and far right and merging them into a consensus.

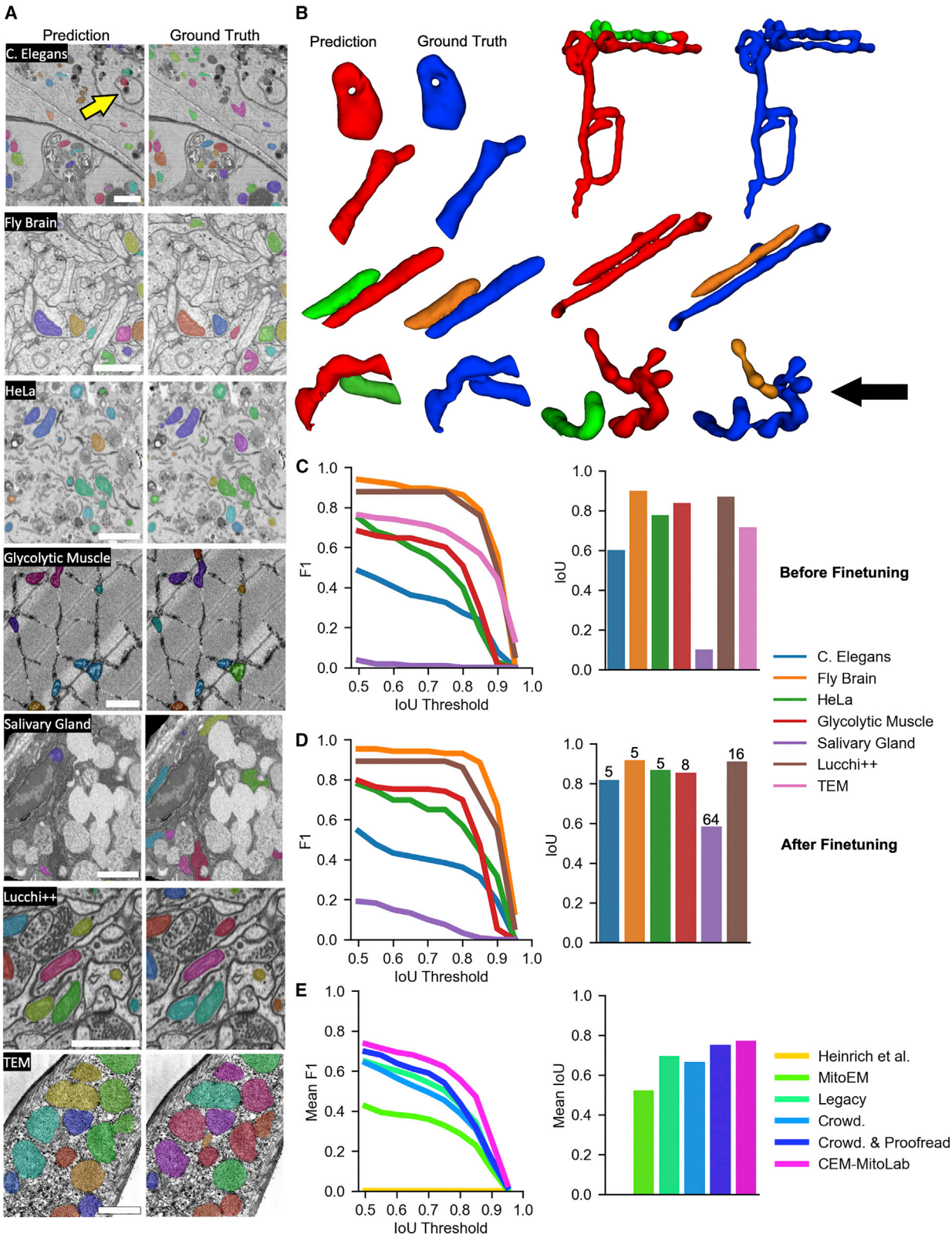


Figure 4. MitoNet results on benchmarks

(A) Representative 2D images showing MitoNet segmentation performance; the left column shows predictions and right column shows ground truth. Top to bottom: *C. elegans*, fly brain, HeLa cell, glycolytic muscle, salivary gland, Lucchi++, and TEM benchmarks. Yellow arrow, representative false positive.

(legend continued on next page)

Table 1. MitoNet performance on the benchmarks

Dataset	IoU	F1@50	F1@75	AP@50	AP@75	PQ
C. elegans	0.604	0.483	0.325	0.318	0.194	0.376
Fly brain	0.902	0.940	0.885	0.887	0.794	0.831
HeLa	0.791	0.728	0.503	0.573	0.336	0.571
Glycolytic muscle	0.840	0.682	0.601	0.518	0.430	0.557
Salivary gland	0.103	0.036	0.008	0.018	0.004	0.022
Lucchi++	0.871	0.879	0.879	0.784	0.784	0.783
TEM	0.717	0.763	0.682	0.654	0.559	0.670

All metrics are calculated in 3D excluding those for the TEM benchmark. Metrics for the TEM benchmark are averages (N = 100). AP, average precision, PQ, panoptic quality.

immediately useful for visualizing the data and identifying dramatically different mitochondrial morphologies. The 2D slice and full 3D instance segmentations of mitochondria from full ROIs are shown in Figure 5A, whereas mitochondria from individual cells, and representative instances of mitochondria showing dramatic variation in morphologies, are shown in Figure 5B. Mitochondria in both kidney volumes were densely packed into rows, polarized toward basolateral surfaces, and revealed to have flattened morphologies. Some 6% of mitochondria in the proximal tubule (middle row) were fused into large and complex networks in sharp contrast to the much less branched morphologies in the distal tubule. Thus, with automatic MitoNet segmentation, limited morphological assessments from light microscopy imaging of proximal tubules[54] can now be extended and quantified at an individual organelle level with the resolution of volume EM. Mitochondria in the hepatocytes appeared to be relatively simple tubes and spheroids in line with previous analyses[57–60]; however, many instances were visible with long and skinny protrusions—possibly nanotunnels[61]—extending into the surrounding cytoplasm. As further evidence of the model's ability to generalize, we observed the accurate segmentations of mitochondria in endothelial cells incidentally captured within the volumes as well (Video S1).

Notably, 89%, 80%, and 60% of randomly sampled instances from the hepatocyte, distal tubule, and proximal tubule volumes, respectively, required no edits. In line with our results on the benchmarks, MitoNet predictions rarely required per-pixel corrections, but merge and split errors occurred in a data-specific manner (Figure 5C). For example, 12% of sampled distal tubule instances required splitting, whereas 21% in the proximal tubules needed merging—multiple merge operations were often required for complex mitochondria. With empanada's point-and-click proofreading operations (Figure S7) we took under 2 h to split and merge all the sampled distal tubule instances. Proofread instance segmentations can be used for a variety of quantitative analyses. We assessed mitochondrial polarization by measuring the distribution of their distances from basolateral surfaces in distal and proximal tubules (Figure 5D) and discovered statistically significant differences (median distance in the distal tubule, 3.38 μm; proximal tubule, 0.86 μm, $p < 0.001$). Additionally, other simple measurements accessible only with instance segmentation, e.g., volume, surface area, cross-sectional radius, elongation, and flatness, also showed intriguing differences between these groups (Figure 5D, statistical measures summarized in Table S7). Such large-scale observations of mitochondrial architectures in 3D and at high resolutions may reveal new and unexpected insights into kidney and other tissue pathologies.

DISCUSSION

In this paper, we have reported several resources for the growing volume EM community: a massive and heterogeneous unlabeled cellular EM image dataset that samples over 2 PB of EM data (CEM1.5M), a similarly heterogeneous dataset of >135,000 instances of labeled mitochondria (CEM-MitoLab), a model trained on this dataset (MitoNet), a set of diverse of 2D and 3D benchmarks (MitoNet Benchmarks) for model testing, and a Python package and napari plugin (empanada) for immediate use of MitoNet with proofreading tools. We also describe these datasets with a simple spreadsheet implementation of recommended metadata for biological images (REMBI)[62] containing image and biological metadata. Together, these resources relieve the "segmentation bottleneck," a current constraint in volume EM workflows otherwise replete with technological and application advances. We note that CEM1.5M and empanada can be used for other features in cellular EM, and although CEM-MitoLab, MitoNet, and benchmarks are mitochondria-specific resources, their performance validates the strategy of broad but shallow sampling of cellular contexts to create general segmentation models for any organelle. The heterogeneity of training data is crucial, as shown by the result that models trained on our noisy crowdsourced annotations generalize better than models trained on narrow but expertly labeled data like MitoEM. Expert proofreading was still necessary to achieve our best results, so broader expert participation in future shared efforts could greatly benefit the field. Similarly, homogeneous volume EM benchmarks derived from connectomics data are poor tests of generalization. The Lucchi++ benchmark has been mined to the point of performance saturation.[51,63] The assortment of mitochondria, contexts, quality, and overall difficulty in our benchmarks offer a stiff test for new models, but future benchmarks must be continually expanded to avoid exhaustion.

We made an important decision to forego 3D approaches. This allowed us to leverage abundant 2D datasets to represent a much broader range of cells and tissues and also maximized images sampled—our crowdsourced dataset of ~6,000 2D labeled images would have equated to just 30 3D labeled images of

(B) Representative 3D ground truth and predicted segmentations from MitoNet on the volume EM benchmarks. Red and green, predicted mitochondrial instances; blue and orange, ground truth instances. Black arrow, example of segmentation expected to return a high IoU but low F1 score.
(C) Left, MitoNet F1 score on each benchmark as a function of IoU threshold; right, IoU scores.
(D) Left, MitoNet F1 scores on volume EM benchmarks as a function of IoU threshold, after model fine-tuning on a small fraction of labeled patches; right, IoU scores achieved by the fine-tuned models. Numbers indicate the number of patches used for fine-tuning.
(E) Left, comparison of mean F1 score for models trained on different datasets plotted against IoU threshold; right, mean IoU scores. All benchmarks except the salivary gland are included in the mean. Crowd., crowdsourced. Scale bars, 1 μm.

Table 2. Average results on all benchmarks (excluding the salivary gland) by models pretrained on CEM1.5M and trained on different labeled datasets

Training dataset	Mean IoU	Mean F1@50	Mean F1@75	Mean AP@50	Mean AP@75	Mean PQ
Heinrich et al.[10]	0.001	0.001	0.001	0.001	0.000	0.001
MitoEM	0.524	0.425	0.328	0.339	0.266	0.346
Legacy	0.697	0.648	0.504	0.520	0.384	0.530
Crowdsourced	0.667	0.642	0.455	0.523	0.358	0.512
Crowdsourced and proofread	0.753	0.698	0.542	0.568	0.411	0.566
CEM-MitoLab	**0.775**	**0.737**	**0.622**	**0.623**	**0.504**	**0.621**

comparable dimensions. Staying in the 2D regime keeps hardware requirements low as well, congruent with our belief that true democratization demands that tools be usable without DL expertise or expensive compute resources—although, we note that our workflows work both on consumer-grade laptops with no GPUs and on high performance computing (HPC) systems with many. The methods we developed in this work, such as median filtering and ortho-plane inference, retain the efficiencies of labeling and running model inference in 2D but utilize 3D context where possible. As shown by our results with the UNet-BC model, post processing with full 3D context can improve instance segmentation at the cost of usability. We expect native 3D model architectures to surpass our methods eventually. Even then, empanada can be used to quickly segment and proofread the 3D data needed to train such models.

True human-level mitochondrial instance segmentation on any 2D or 3D EM image may be possible. Better methods are needed to segment small and closely packed instances, and unusual cellular contexts or low-contrast and resolution images pose some difficulties. That said, semantic segmentations were consistently accurate on most datasets tested, and a proof-of-principle experiment on kidney and liver volumes shows that MitoNet is already a powerful and accessible tool for rapid visualization and accurate quantification of mitochondrial morphologies, even without proofreading for some datasets (Table S7). Current DL pipelines using specialist models rely on repeated cycles of human-in-the-loop annotation, training, inference, and proofreading. With generalist DL segmentation models, like MitoNet for mitochondria, this cycle is reduced to just inference and proofreading on any EM dataset. Even for datasets where out-of-the-box inference is poor, our napari plugin, empanada, enables quick fine-tuning on sparsely sampled 2D patches for improved results. Future expansions of the CEM resources will further enable model generality irrespective of feature segmented. Such models, whether for semantic, instance, or panoptic segmentation, can be trained using empanada and broadly shared for deployment.

Volume EM is catalyzing a "quiet revolution"[64] by enabling large-volume high-resolution 3D images of cellular and subcellular features, revealing insights into connectomics and cell biology. Accurate and efficient extraction of features of interest from these massive ultrastructural images lends itself to computational solutions. Here, we create large-scale and relevant training data resources and train a generalist model to segment mitochondria in any EM image. We also release an easy-to-use software plugin for users to deploy this model on their own data and execute downstream "cleanup" steps to generate polished mitochondrial segmentations. In tandem with recent initiatives to generate and share large-volume EM reconstructions,[36] our methods provide a blueprint for future applications, thereby expanding the segmentation toolkit and helping accelerate discoveries in this exciting field.

$$w_i = \frac{n_i^{-\gamma}}{\sum_{j=1}^{N} w_i}$$

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead Contact
  - Materials Availability
  - Data and Code Availability
- METHOD DETAILS
  - CEM1.5M Creation
  - Crowdsourced Annotation
  - Labeled and benchmark dataset creation
  - CEM1.5M Pre-training
  - MitoNet model architecture, training parameters, and post-processing
  - Unet-BC model architecture, training parameters, and post-processing
  - Benchmark inference and evaluation
  - MitoNet finetuning
  - OpenOrganelle volume inference and analysis
- QUANTIFICATION AND STATISTICAL ANALYSIS
- ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cels.2022.12.006.

ACKNOWLEDGMENTS

This project has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, under contract no. 75N91019D00024. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services nor does mention of trade names, commercial products, or organizations imply

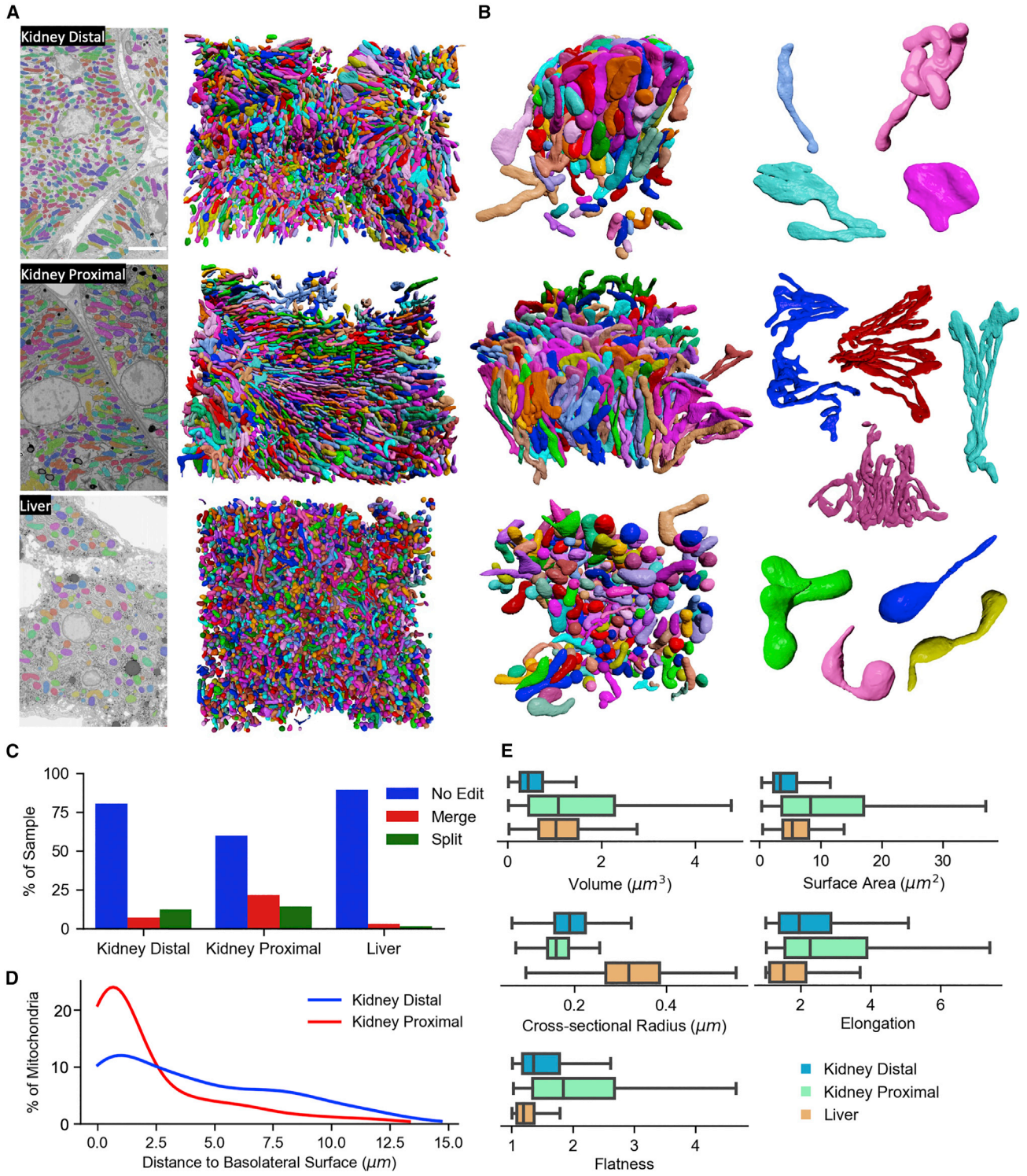


Figure 5. MitoNet results on volumes of mouse liver and kidney

(A) Rows from top to bottom correspond to the kidney distal tubule, kidney proximal tubule, and liver. The left column shows representative 2D images of MitoNet segmentation (scale bars, 5 μm); the right column shows 3D predictions on the entire volume (small and boundary objects removed).

(B) Left column shows a zoomed-in ROI of raw model predictions (basolateral surfaces of cells on top); the right column shows representative mitochondrial models after manual cleanup.

(C) Plot of the fraction and type of cleanup operation required for a randomly chosen sample of model-predicted instances from kidney distal (n = 347), kidney proximal (n = 256), and liver (n = 319) tissue.

(legend continued on next page)

endorsement by the U.S. Government. This publication uses data generated via the https://www.zooniverse.org/ platform, the development of which is funded by generous support, including a Global Impact Award from Google and a grant from the Alfred P. Sloan Foundation. We thank Helen Spiers, Martin Jones, and Lucy Collinson for their help with Zooniverse. We thank the WHK-SIP program and batch of 2021 student annotators, especially student leaders Ella Fitzgerald and Taeeun Kim. Valentina Baena, Adam Harned, Kunio Nagashima, and Heather Berensmann helped perform the proofreading. Patrick Friday helped curate metadata, and Aayush Bhatawadekar contributed to renderings and plugin development. Finally, we appreciate the volume EM community, especially the Data Working Group, for general discussions. This work utilized the computational resources of the NIH HPC Biowulf cluster. (http://hpc.nih.gov).

AUTHOR CONTRIBUTIONS

R.C. wrote the software and devised the algorithms. R.C. and K.N. conceived of the project and wrote the paper.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Peddie, C.J., and Collinson, L.M. (2014). Exploring the third dimension: volume electron microscopy comes of age. Micron *61*, 9–19. https://doi.org/10.1016/J.MICRON.2014.01.009.
2. Titze, B., and Genoud, C. (2016). Volume scanning electron microscopy for imaging biological ultrastructure. Biol. Cell *108*, 307–323. https://doi.org/10.1111/BOC.201600024.
3. Scheffer, L.K., Xu, C.S., Januszewski, M., Lu, Z., Takemura, S.Y., Hayworth, K.J., Huang, G.B., Shinomiya, K., Maitlin-Shepard, J., Berg, S., et al. (2020). A connectome and analysis of the adult drosophila central brain. eLife *9*, 1–74. https://doi.org/10.7554/eLife.57443.
4. Turner, N.L., Macrina, T., Bae, J.A., Yang, R., Wilson, A.M., Schneider-Mizell, C., Lee, K., Lu, R., Wu, J., Bodor, A.L., et al. (2022). Reconstruction of neocortex: organelles, compartments, cells, circuits, and activity. Cell *185*. 1082–1100.e24. https://doi.org/10.1016/J.CELL.2022.01.023.
5. Yin, W., Brittain, D., Borseth, J., Scott, M.E., Williams, D., Perkins, J., Own, C.S., Murfitt, M., Torres, R.M., Kapner, D., et al. (2020). A petascale automated imaging pipeline for mapping neuronal circuits with high-throughput transmission electron microscopy. Nat. Commun. *11*, 4949. https://doi.org/10.1038/s41467-020-18659-3.
6. Januszewski, M., Kornfeld, J., Li, P.H., Pope, A., Blakely, T., Lindsey, L., Maitin-Shepard, J., Tyka, M., Denk, W., and Jain, V. (2018). High-precision automated reconstruction of neurons with flood-filling networks. Nat. Methods *15*, 605–610. https://doi.org/10.1038/s41592-018-0049-4.
7. Berning, M., Boergens, K.M., and Helmstaedter, M. (2015). SegEM: efficient image analysis for high-resolution connectomics. Neuron *87*, 1193–1206. https://doi.org/10.1016/j.neuron.2015.09.003.
8. Dorkenwald, S., Schubert, P.J., Killinger, M.F., Urban, G., Mikula, S., Svara, F., and Kornfeld, J. (2017). Automated synaptic connectivity inference for volume electron microscopy. Nat. Methods *14*, 435–442. https://doi.org/10.1038/nmeth.4206.
9. Funke, J., Tschopp, F., Grisaitis, W., Sheridan, A., Singh, C., Saalfeld, S., and Turaga, S.C. (2019). Large scale image segmentation with structured loss based deep learning for connectome reconstruction. IEEE Trans. Pattern Anal. Mach. Intell. *41*, 1669–1680. https://doi.org/10.1109/TPAMI.2018.2835450.
10. Heinrich, L., Bennett, D., Ackerman, D., Park, W., Bogovic, J., Eckstein, N., Petruncio, A., Clements, J., Pang, S., Xu, C.S., et al. (2021). Whole-cell organelle segmentation in volume electron microscopy. Nature *599*, 141–146. https://doi.org/10.1038/s41586-021-03977-3.
11. Žerovnik Mekuč, M., Bohak, C., Hudoklin, S., Kim, B.H., Romih, R., Kim, M.Y., and Marolt, M. (2020). Automatic segmentation of mitochondria and endolysosomes in volumetric electron microscopy data. Comput. Biol. Med. *119*, 103693. https://doi.org/10.1016/j.compbiomed.2020.103693.
12. Liu, J., Li, L., Yang, Y., Hong, B., Chen, X., Xie, Q., and Han, H. (2020). Automatic reconstruction of mitochondria and endoplasmic reticulum in electron microscopy volumes by deep learning. Front. Neurosci. *14*, 599. https://doi.org/10.3389/fnins.2020.00599.
13. Spiers, H., Songhurst, H., Nightingale, L., de Folter, J., Zooniverse Volunteer Community, Hutchings, R., Peddie, C.J., Weston, A., Strange, A., Hindmarsh, S., et al. (2021). Deep learning for automatic segmentation of the nuclear envelope in electron microscopy data, trained with volunteer segmentations. Traffic *22*, 240–253. https://doi.org/10.1111/TRA.12789.
14. Guay, M.D., Emam, Z.A.S., Anderson, A.B., Aronova, M.A., Pokrovskaya, I.D., Storrie, B., and Leapman, R.D. (2021). Dense cellular segmentation for EM using 2D–3D neural network ensembles. Sci. Rep. *11*, 2561. https://doi.org/10.1038/s41598-021-81590-0.
15. Wei, D., Lin, Z., Franco-Barranco, D., Wendt, N., Liu, X., Yin, W., Huang, X., Gupta, A., Jang, W.D., Wang, X., et al. (2020). MitoEM dataset: large-scale 3D mitochondria instance segmentation from EM images. Med. Image Comput. Comput. Assist. Interv. *12265*, 66–76. https://doi.org/10.1007/978-3-030-59722-1_7.
16. Müller, A., Schmidt, D., Xu, C.S., Pang, S., D'Costa, J.V., Kretschmar, S., Münster, C., Kurth, T., Jug, F., Weigert, M., et al. (2021). 3D FIB-SEM reconstruction of microtubule–organelle interaction in whole primary mouse β cells. J. Cell Biol. *220*, e202010039. https://doi.org/10.1083/JCB.202010039.
17. Buhmann, J., Sheridan, A., Malin-Mayor, C., Schlegel, P., Gerhard, S., Kazimiers, T., Krause, R., Nguyen, T.M., Heinrich, L., Lee, W.A., et al. (2021). Automatic detection of synaptic partners in a whole-brain Drosophila electron microscopy data set. Nat. Methods *18*, 771–774. https://doi.org/10.1038/s41592-021-01183-7.
18. Meyer, J.N., Leuthner, T.C., and Luz, A.L. (2017). Mitochondrial fusion, fission, and mitochondrial toxicity. Toxicology *391*, 42–53. https://doi.org/10.1016/j.tox.2017.07.019.
19. Zhang, L., Trushin, S., Christensen, T.A., Bachmeier, B.V., Gateno, B., Schroeder, A., Yao, J., Itoh, K., Sesaki, H., Poon, W.W., et al. (2016). Altered brain energetics induces mitochondrial fission arrest in Alzheimer's disease. Sci. Rep. *6*, 18725. https://doi.org/10.1038/srep18725.
20. Pernas, L., and Scorrano, L. (2016). Mito-morphosis: mitochondrial fusion, fission, and cristae remodeling as key mediators of cellular function. Annu. Rev. Physiol. *78*, 505–531. https://doi.org/10.1146/annurev-physiol-021115-105011.
21. Delgado, T., Petralia, R.S., Freeman, D.W., Sedlacek, M., Wang, Y.X., Brenowitz, S.D., Sheu, S.H., Gu, J.W., Kapogiannis, D., Mattson, M.P., et al. (2019). Comparing 3D ultrastructure of presynaptic and postsynaptic mitochondria. Biol. Open *8*, bio044834. https://doi.org/10.1242/bio.044834.
22. Glancy, B., Kim, Y., Katti, P., and Willingham, T.B. (2020). The functional impact of mitochondrial structure across subcellular scales. Front. Physiol. *11*, 541040. https://doi.org/10.3389/fphys.2020.541040.

(D) Plot of distance to the nearest basolateral surface, in microns, for randomly sampled mitochondria from kidney distal (blue) and kidney proximal (red) volumes after cleanup.
(E) Boxplot comparisons of mitochondrial volume, surface area, cross-sectional radius, elongation, and flatness across the three volumes after cleanup (outliers not shown). Blue, kidney distal (n = 405 for D and E); green, kidney proximal (n = 250); orange, liver (n = 321).

23. Nunnari, J., and Suomalainen, A. (2012). Mitochondria: in sickness and in health. Cell *148*, 1145–1159. https://doi.org/10.1016/J.CELL.2012.02.035.

24. Bhargava, P., and Schnellmann, R.G. (2017). Mitochondrial energetics in the kidney. Nat. Rev. Nephrol. *13*, 629–646. https://doi.org/10.1038/NRNEPH.2017.107.

25. Doke, T., and Susztak, K. (2022). The multifaceted role of kidney tubule mitochondrial dysfunction in kidney disease development. Trends Cell Biol. *32*, 841–853. https://doi.org/10.1016/J.TCB.2022.03.012.

26. Emma, F., Montini, G., Parikh, S.M., and Salviati, L. (2016). Mitochondrial dysfunction in inherited renal disease and acute kidney injury. Nat. Rev. Nephrol. *12*, 267–280. https://doi.org/10.1038/NRNEPH.2015.214.

27. Leonard, A.P., Cameron, R.B., Speiser, J.L., Wolf, B.J., Peterson, Y.K., Schnellmann, R.G., Beeson, C.C., and Rohrer, B. (2015). Quantitative analysis of mitochondrial morphology and membrane potential in living cells using high-content imaging, machine learning, and morphological binning. Biochim. Biophys. Acta *1853*, 348–360. https://doi.org/10.1016/j.bbamcr.2014.11.002.

28. Nikolaisen, J., Nilsson, L.I.H., Pettersen, I.K.N., Willems, P.H.G.M., Lorens, J.B., Koopman, W.J.H., and Tronstad, K.J. (2014). Automated quantification and integrative analysis of 2D and 3D mitochondrial shape and network properties. PLoS One *9*, e101365. https://doi.org/10.1371/journal.pone.0101365.

29. Talwar, A., Lin, Z., Wei, D., Wu, Y., Zheng, B., Zhao, J., Jang, W.D., Wang, X., Lichtman, J., and Pfister, H. (2020). A topological nomenclature for 3D shape analysis in connectomics. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 4245–4253.

30. Miyazono, Y., Hirashima, S., Ishihara, N., Kusukawa, J., Nakamura, K.-I., and Ohta, K. (2018). Uncoupled mitochondria quickly shorten along their long axis to form indented spheroids, instead of rings, in a fission-independent manner. Sci. Rep. *8*, 350. https://doi.org/10.1038/s41598-017-18582-6.

31. Vincent, A.E., White, K., Davey, T., Taylor, R.W., Turnbull, D.M., and Picard, M. (2019). Quantitative 3D mapping of the human skeletal muscle mitochondrial network. Cell Rep. *26*. 996–1009.e4. https://doi.org/10.1016/j.celrep.2019.01.010.

32. Bleck, C.K.E., Kim, Y., Willingham, T.B., and Glancy, B. (2018). Subcellular connectomic analyses of energy networks in striated muscle. Nat. Commun. *9*, 5111. https://doi.org/10.1038/s41467-018-07676-y.

33. Abrisch, R.G., Gumbin, S.C., Wisniewski, B.T., Lackner, L.L., and Voeltz, G.K. (2020). Fission and fusion machineries converge at ER contact sites to regulate mitochondrial morphology. J. Cell Biol. *219*, e201911122. https://doi.org/10.1083/JCB.201911122.

34. Tamada, H., Kiryu-Seo, S., Hosokawa, H., Ohta, K., Ishihara, N., Nomura, M., Mihara, K., Nakamura, K.I., and Kiyama, H. (2017). Three-dimensional analysis of somatic mitochondrial dynamics in fission-deficient injured motor neurons using FIB/SEM. J. Comp. Neurol. *525*, 2535–2548. https://doi.org/10.1002/cne.24213.

35. Conrad, R., and Narayan, K. (2021). CEM500K, a large-scale heterogeneous unlabeled cellular electron microscopy image dataset for deep learning. eLife *10*. https://doi.org/10.7554/eLife.65894.

36. Xu, C.S., Pang, S., Shtengel, G., Müller, A., Ritter, A.T., Hoffman, H.K., Takemura, S.-Y., Lu, Z., Pasolli, H.A., Iyer, N., et al. (2021). An open-access volume electron microscopy atlas of whole cells and tissues. Nature *599*, 147–151. https://doi.org/10.1038/s41586-021-03992-4.

37. Casser, V., Kang, K., Pfister, H., and Haehn, D. (2020). Fast mitochondria detection for connectomics. Proceedings of the Machine Learning Research *121*, 111–120.

38. Lucchi, A., Li, Y., and Fua, P. (2013). Learning for structured prediction using approximate subgradient descent with working sets. 2013 IEEE Conference on Computer Vision and Pattern Recognition, 1987–1994. https://doi.org/10.1109/CVPR.2013.259.

39. Riddle, D.L., Blumenthal, T., Meyer, B.J., and Priess, J.R. (1997). *C. elegans* II (Cold Spring Harbor Laboratory Press).

40. Cheng, B., Collins, M.D., Zhu, Y., Liu, T., Huang, T.S., Adam, H., and Chen, L.-C. (2020). Panoptic-DeepLab: a simple, strong, and fast baseline for bottom-up panoptic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

41. Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (Springer Verlag), pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28.

42. He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2020). Mask R-CNN. IEEE Transactions on Pattern Analysis and Machine Intelligence *42*, 386–397. https://doi.org/10.1109/TPAMI.2018.2844175.

43. Cheng, B., Schwing, A.G., and Kirillov, A. (2021). Per-pixel classification is not all you need for semantic segmentation. In 35th Conference on Neural Information Processing Systems (NeurIPS 2021).

44. Kirillov, A., Wu, Y., He, K., and Girshick, R. (2020). PointRend: image segmentation as rendering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9799–9808. https://doi.org/10.48550/arXiv.1912.08193.

45. Kuhn, H.W. (1955). The Hungarian method for the assignment problem. Nav. Res. Logist. Q. *2*, 83–97. https://doi.org/10.1002/NAV.3800020109.

46. Conrad, R., Lee, H., and Narayan, K. (2020). Enforcing prediction consistency across orthogonal planes significantly improves segmentation of FIB-SEM image volumes by 2D neural networks. Microsc. Microanal. *26*, 2128–2130. https://doi.org/10.1017/S143192762002053X.

47. Tan, M., Pang, R., and Le, Q.V. (2020). EfficientDet: scalable and efficient object detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10778–10787.

48. Lin, Z., Wei, D., Lichtman, J., and Pfister, H. (2021). PyTorch connectomics: a scalable and flexible segmentation framework for EM connectomics. https://doi.org/10.48550/arxiv.2112.05754.

49. He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (IEEE Computer Society), pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.

50. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. https://doi.org/10.48550/arxiv.2006.09882.

51. Xiao, C., Chen, X., Li, W., Li, L., Wang, L., Xie, Q., and Han, H. (2018). Automatic mitochondria segmentation for EM data using a 3D supervised convolutional network. Front. Neuroanat. *12*, 92. https://doi.org/10.3389/fnana.2018.00092.

52. Manoli, I., Sysol, J.R., Li, L., Houillier, P., Garone, C., Wang, C., Zerfas, P.M., Cusmano-Ozog, K., Young, S., Trivedi, N.S., et al. (2013). Targeting proximal tubule mitochondrial dysfunction attenuates the renal disease of methylmalonic acidemia. Proc. Natl. Acad. Sci. USA *110*, 13552–13557. https://doi.org/10.1073/pnas.1302764110.

53. Bergeron, M., Guerette, D., Forget, J., and Thiery, G. (1980). Three-dimensional characteristics of the mitochondria of the rat nephron. Kidney Int. *17*, 175–185. https://doi.org/10.1038/ki.1980.21.

54. Taguchi, K., Elias, B.C., Krystofiak, E., Qian, S., Sant, S., Yang, H., Fogo, A.B., and Brooks, C.R. (2021). Quantitative super-resolution microscopy reveals promoting mitochondrial interconnectivity protects against AKI. Kidney360 *2*, 1892–1907. https://doi.org/10.34067/KID.0001602021.

55. Ghazi, S., Bourgeois, S., Gomariz, A., Bugarski, M., Haenni, D., Martins, J.R., Nombela-Arrieta, C., Unwin, R.J., Wagner, C.A., Hall, A.M., and Craigie, E. (2020). Multiparametric imaging reveals that mitochondria-rich intercalated cells in the kidney collecting duct have a very high glycolytic capacity. FASEB J. *34*, 8510–8525. https://doi.org/10.1096/fj.202000273R.

56. Hall, A.M., Unwin, R.J., Parker, N., and Duchen, M.R. (2009). Multiphoton imaging reveals differences in mitochondrial function between nephron

segments. J. Am. Soc. Nephrol. *20*, 1293–1302. https://doi.org/10.1681/ASN.2008070759.

57. Parlakgül, G., Arruda, A.P., Pang, S., Cagampan, E., Min, N., Güney, E., Lee, G.Y., Inouye, K., Hess, H.F., Xu, C.S., and Hotamışlıgil, G.S. (2022). Regulation of liver subcellular architecture controls metabolic homeostasis. Nature *603*, 736–742. https://doi.org/10.1038/S41586-022-04488-5.
58. Chun Chung, G.H.C., Gissen, P., Stefan, C.J., and Burden, J.J. (2022). Three-dimensional characterization of interorganelle contact sites in hepatocytes using serial section electron microscopy. J. Vis. Exp. https://doi.org/10.3791/63496.
59. Kizilyaprak, C., De Bellis, D., Blanchard, W., Daraspe, J., and Humbel, B.M. (2019). FIB-SEM tomography of biological samples: explore the life in 3D. Biological Field Emission Scanning Electron Microscopy, I, 545–566. https://doi.org/10.1002/9781118663233.CH26.
60. Murphy, G.E., Lowekamp, B.C., Zerfas, P.M., Chandler, R.J., Narasimha, R., Venditti, C.P., and Subramaniam, S. (2010). Ion-abrasion scanning electron microscopy reveals distorted liver mitochondrial morphology in murine methylmalonic acidemia. J. Struct. Biol. *171*, 125–132. https://doi.org/10.1016/J.JSB.2010.04.005.
61. Vincent, A.E., Turnbull, D.M., Eisner, V., Hajnóczky, G., and Picard, M. (2017). Mitochondrial nanotunnels. Trends Cell Biol. *27*, 787–799. https://doi.org/10.1016/J.TCB.2017.08.009.
62. Sarkans, U., Chiu, W., Collinson, L., Darrow, M.C., Ellenberg, J., Grunwald, D., Hériché, J.-K., Iudin, A., Martins, G.G., Meehan, T., et al. (2021). REMBI: recommended Metadata for Biological Images—enabling reuse of microscopy data in biology. Nat. Methods *18*, 1418–1422. https://doi.org/10.1038/s41592-021-01166-8.
63. Franco-Barranco, D., Muñoz-Barrutia, A., and Arganda-Carreras, I. (2021). Stable deep neural network architectures for mitochondria segmentation on electron microscopy volumes. Neuroinformatics *20*, 437–450. https://doi.org/10.1007/S12021-021-09556-1/TABLES/6.
64. Narayan, K., and Subramaniam, S. (2015). Focused ion beams in biology. Nat. Methods *12*, 1021–1031. https://doi.org/10.1038/NMETH.3623.
65. Liu, L., Yang, S., Liu, Y., Li, X., Hu, J., Xiao, L., and Xu, T. (2022). DeepContact: high-throughput quantification of membrane contact sites based on electron microscopy imaging. J. Cell Biol. *221*, e202106190. https://doi.org/10.1083/JCB.202106190.
66. Kasthuri, N., Hayworth, K.J., Berger, D.R., Schalek, R.L., Conchello, J.A., Knowles-Barley, S., Lee, D., Vázquez-Reina, A., Kaynig, V., Jones, T.R., et al. (2015). Saturated reconstruction of a volume of neocortex. Cell *162*, 648–661. https://doi.org/10.1016/j.cell.2015.06.054.
67. Perez, A.J., Seyedhosseini, M., Deerinck, T.J., Bushong, E.A., Panda, S., Tasdizen, T., and Ellisman, M.H. (2014). A workflow for the automatic segmentation of organelles in electron microscopy image stacks. Front. Neuroanat. *8*, 126. https://doi.org/10.3389/fnana.2014.00126.
68. Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., and Girshick, R. (2019). Detectron2. https://www.bibsonomy.org/bibtex/2937080a00feedf57847ed214469fa3e7/s352021.
69. Smith, L.N. (2018). A disciplined approach to neural network hyper-parameters: part 1 – learning rate, batch size, momentum, and weight decay. https://doi.org/10.48550/arXiv.1803.09820.
70. Loshchilov, I., and Hutter, F. (2017). Decoupled weight decay regularization. https://doi.org/10.48550/arXiv.1711.05101.
71. Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T., Cubuk, E.D., Le, Q.V., and Zoph, B. (2021). Simple copy-paste is a strong data augmentation method for instance segmentation. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (IEEE Computer Society), pp. 2917–2927. https://doi.org/10.1109/3DV.2016.79.
72. Milletari, F., Navab, N., and Ahmadi, S.A. (2016). V-net: fully convolutional neural networks for volumetric medical image segmentation. 2016 Fourth International Conference on 3D Vision (3DV) *2016*, 565–571. https://doi.org/10.48550/arxiv.1606.04797.
73. McCormick, M., Liu, X., Jomier, J., Marion, C., and Ibanez, L. (2014). ITK: enabling reproducible research and open science. Front. Neuroinform. *8*, 13. https://doi.org/10.3389/fninf.2014.00013.
74. Bernardini, F., Mittleman, J., Rushmeier, H., Silva, C., and Taubin, G. (1999). The ball-pivoting algorithm for surface reconstruction. IEEE Transactions on Visualization and Computer Graphics *5*, 349–359. https://doi.org/10.1109/2945.817351.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
CEM1.5M	This paper	10.6019/EMPIAR-11035
CEM-MitoLab	This paper	10.6019/EMPIAR-11037
Benchmark datasets	This paper	10.6019/EMPIAR-10982
Heinrich et al.[10]	Heinrich et al.[10]	https://open.quiltdata.com/b/janelia-cosem-publications/tree/heinrich-2021a/ (no DOI available)
MitoEM	Wei et al.[15]	https://mitoem.grand-challenge.org/MitoEM/ (no DOI available)
OpenOrganelle mouse liver	OpenOrganelle	10.25378/janelia.16913047.v1
OpenOrganelle mouse kidney	OpenOrganelle	10.25378/janelia.16913035.v1
DeepContact Training Data	Liu et al.[65]	10.6084/m9.figshare.19898404.v3
Software and algorithms		
empanada	This paper	10.5281/zenodo.7373984
empanada-napari	This paper	10.5281/zenodo.7374003
napari	napari contributors	https://doi.org/10.5281/zenodo.3555620
arivis Vision4D	Zeiss	Vision4D

RESOURCE AVAILABILITY

Lead Contact

Further information and requests for resources should be directed to Kedar Narayan (narayank@mail.nih.gov).

Materials Availability

This study did not generate new materials.

Data and Code Availability

Datasets have been deposited at EMPIAR and are publicly available as of the date of publication: CEM1.5M (https://www.ebi.ac.uk/empiar/EMPIAR-11035/), CEM-MitoLab (https://www.ebi.ac.uk/empiar/EMPIAR-11037/), MitoNet Benchmarks (https://www.ebi.ac.uk/empiar/EMPIAR-10982/).

Code for the empanada library is available at https://doi.org/10.5281/zenodo.3555620. Code for the empanada-napari plugin is available at https://doi.org/10.5281/zenodo.7374004. Pre-trained models are available on Zenodo: CEM1.5M weights (https://zenodo.org/record/6453160#.YnPjEy-cbTR), MitoNet (https://zenodo.org/record/6327742#.YnLciy-cbTS).

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

CEM1.5M Creation

The expansion of CEM500K to create the CEM1.5M dataset followed the data standardization and curation protocols presented in our previous work.[35] External datasets were either downloaded in their entirety or, for large datasets stored online in next generation file formats (n5 or zarr), accessed either with the CloudVolume or fibsem-tools APIs. Datasets larger than 5 GB were randomly cropped into 320 cubes of 256^3. Metadata and proper attribution for each dataset is available in File S1. TEM and STEM images were grouped into directories based on imaging project or publication before processing. 3,000 NCI TEM images with magnifications between 1000x (~16 nm pixel) and 3000x (~6 nm pixel) and excluding negative stain and immunogold label images were randomly selected from a library of over $2x10^5$ images. Metadata for these images was unavailable.

Crowdsourced Annotation

Our Zooniverse workflow closely approximated the Etch-a-cell project.[13] Flipbooks of image patches from 3D data (8-bit tif stacks of five consecutive images of 224x224) were contrast rescaled to 25-235 and interpolated to 480x480 for easier viewing. For 2D images contrast was similarly rescale but crops of 512x512 were used. Over the course of the project, the Zooniverse chat channel and a formal instruction session were used to review examples and explain errors. During expert proofreading, the correction of FP and FN detections was emphasized over pixel-level painting; disagreements were discussed and resolved as a group. All annotations were passed through a connected components filter after proofreading.

The input to the consensus algorithm was a set of **N** annotations (the retirement limit) containing **K** mitochondrial detections (Figure S3i). An undirected graph was initialized where each node corresponded to one of the **K** detections. Edges were added to the graph to connect detections with mask IoU scores that exceeded a small threshold value of 0.1 (the same algorithm, when applied during ortho-plane inference, used a threshold of 0.01). Each connected component in the detection graph was then processed independently. Nodes in a connected component were organized into cliques where all detections within a clique shared edges with IoU scores greater than 0.75 (Figure S3ii, iii). Edges between detections in different cliques were retained. An iterative algorithm was applied to determine whether cliques should be merged or remain split. First, the most connected clique (i.e., the one with the most edges; node C in Figures S3ii and S3iv; node A in Figures S3iii and S3v) was selected. Second, if any of the most connected clique's neighbors contained more detections, then the most connected clique was dissolved and its detections were pushed out to each of its neighbors (C in Figure S3iv). Otherwise, all the most connected clique's neighbors were dissolved and their detections were merged into the most connected clique (B and D in Figure S3iv; B and C into A in Figure S3v). These two steps were repeated until no edges were left between cliques. Each clique represented an object instance. The detection masks within a clique were added together to form an image where each pixel had a value from 1 to **N**. A vote threshold was applied to create the final binary instance mask. All instances were combined into the final consensus annotation (Figure S3vi).

Labeled and benchmark dataset creation

The "legacy" portion of CEM-MitoLab came from publicly available mitochondrial segmentations (Kasthuri++, Guay, UroCell, MitoEM, Heinrich et al., and Perez et al.)[10,37,66,67] and previous in-house segmentation projects. Of these, MitoEM already had instance labels, and a connected components filter was applied to convert other semantic labels to instance segmentations. All in-house segmentation projects were proofread or annotated by experienced researchers. The projects included annotations of whole volumes, TEM images, and a random selection of patches from CEM500K[35]; none were derived from the benchmark datasets. For volumes and TEM images, 512x512 crops were taken from xy planes, or all three planes for isotropic volumes, and passed through the deduplication and filtering pipeline to remove redundant and uninformative patches. Legacy images were used as-is except for MitoEM-H and MitoEM-R volumes which were binned to 16 nm pixels and Heinrich et al.[10] images which were downloaded at 8 nm resolution. To create the alternative training datasets in Figure 4, MitoEM and Heinrich et al.[10] ground truth ROIs were sliced to 2D images. For MitoEM, patches of size 512x512 were cropped from xy planes only. The Heinrich et al.[10] volumes were cropped into patches of 224x224 or smaller from xy, xz and yz planes, and 2D segmentations were passed through a connected components filter.

The HeLa cell, C. elegans, fly brain and salivary gland benchmark datasets were annotated by ariadne (https://ariadne.ai/). The glycolytic muscle[32] dataset was annotated in-house. The benchmark derived from Lucchi++ was generated by applying a connected components filter to the binary mitochondrial labelmaps followed by manual proofreading. To be rigorous, we excluded from CEM1.5M and CEM-MitoLab not just the benchmark datasets, but also images related to them, e.g., the entire Lucchi++ training dataset, all other drosophila neuronal data from OpenOrganelle, and FIB-SEM and STEM data from the in-house C. elegans project.

Images in the TEM benchmark were gathered from internal NCI TEM repositories, the Cell Image Library (http://www.cellimagelibrary.org/home), Nanotomy (http://www.nanotomy.org/), and the WormImage Database (https://www.wormimage.org/), and metadata included where possible. To match the TEM pixel sizes sampled in the CEM-MitoLab dataset, images were minimally downsampled such that none had pixels smaller than 6 nm. Ground truth segmentations were created manually by two experienced researchers.

CEM1.5M Pre-training

Unsupervised pre-training used the SwAV algorithm.[50] A ResNet50[49] model was trained for 200 epochs with a batch size of 256; other hyperparameters used defaults defined in https://github.com/facebookresearch/swav. Image augmentations included 360-degree rotations, randomly resized crops, brightness and contrast jitter, random Gaussian blur and noise, and horizontal and vertical flips. To correct for the imbalance in the number of patches per dataset, weighted random sampling was applied. Weights per dataset were calculated by:

$$w_i = \frac{n_i^{-\gamma}}{\sum_{j=1}^{N} w_i}$$

γ was a float from 0 to 1, n_i was the number of patches from the i^{th} dataset, and N was the total number of datasets. We used γ=0.5.

MitoNet model architecture, training parameters, and post-processing

MitoNet was based on Panoptic-DeepLab (PDL). PDL is an encoder-decoder style network that uses atrous spatial pyramid pooling (ASPP) to integrate features over multiple scales; our configuration included one ASPP decoder for semantic segmentation prediction and another for instance center and offsets prediction. A ResNet50, pre-trained as detailed above, was used as the encoder network. To preserve spatial information, strides in the last downsampling layer were replaced with dilation such that the output from the encoder was 16x smaller than the input. Dilation rates in the ASPP modules were set to 2, 4, and 6 with 256 channels per convolution and dropout probability of 0.5. The semantic and instance decoders had a single level at which the output from the first layer in the encoder was fused to the interpolated ASPP output via depthwise separable convolution. Convolutions had 32 channels in the semantic decoder and 16 in the instance decoder. The semantic segmentation, instance center and offset heads used a single depthwise separable convolution with kernel size of 5. The PointRend[44] module was added to refine the semantic segmentation. Module parameters used the defaults defined in Detectron2.[68] Briefly, during training, the segmentation logits were upsampled by a factor of 4 to the original image resolution and then refined in a single step. During evaluation, two refinement steps were applied with the segmentation logits being upsampled by a factor of 2 at each step. An arbitrary number of interpolation and refinement steps can be applied sequentially to upsample the segmentation to a desired resolution. The overall loss function was:

$$L_{total} = L_{sem} + \alpha L_{center} + \beta L_{offset} + \gamma L_{pointrend}$$

α, β, γ were constants set to 200, 0.01 and 1, respectively. L_{sem} was the semantic segmentation loss computed as the bootstrapped (binary) cross-entropy where only the top 20% of cross-entropy values were averaged across a batch. L_{center} was the instance center regression loss and L_{offset} was the center offset loss calculated as mean squared error and absolute error (L1), respectively. $L_{pointrend}$ was the PointRend loss calculated as the (binary) cross-entropy.

The only difference in architecture between MitoNet and the MitoNet-mini model was the replacement of ASPP modules with BiFPN modules.[47] Both the semantic and instance decoders were replaced with a stack of three BiFPN modules each with a feature dimension of 128. The model architecture and loss function were otherwise the same.

MitoNet and MitoNet-mini were trained on CEM-MitoLab using the One Cycle learning rate policy[69] with AdamW[70] for 120 epochs. The max learning rate was set to 0.003 for MitoNet and 0.001 for MitoNet-mini with weight decay of 0.1 and momentum cycled from 0.85 to 0.95. Learning rate warmup lasted for the first 30% of training epochs. The batch size was 64. Image augmentations included large scale jitter,[71] random cropping of a 256x256 patch, 360-degree rotations, brightness and contrast adjustments, and vertical and horizontal flips. Weighted sampling, as defined above for CEM1.5, used γ=0.3. Postprocessing of the model outputs in an instance segmentation used the method developed for Panoptic-DeepLab. In all experiments, the non-maximum suppression kernel size was set to 7 and the center confidence threshold was set to 0.1.

Unet-BC model architecture, training parameters, and post-processing

For the Unet-BC model we used the implementation provided in PyTorch Connectomics.[48] The model was trained to predict one semantic segmentation for mitochondria and another for mitochondrial contours (i.e., boundaries between mitochondria and background). The only change made to the default model configuration was a doubling of the number of filters in each layer. The overall loss function was:

$$L_{total} = L_{sem} + L_{boundary} + \alpha\left(L_{sem\_dice} + L_{boundary\_dice}\right)$$

α was a constant value of 0.5. L_{sem} and $L_{boundary}$ were the binary cross-entropy losses calculated for the semantic and boundary segmentations respectively. L_{sem_dice} and $L_{boundary\_dice}$ were the dice losses[72] calculated for the semantic and boundary segmentations respectively. The training protocol was identical to that used for MitoNet, CEM1.5M pre-trained weights could not be used to initialize any Unet-BC models because the encoder used was not a ResNet50.

To postprocess the model outputs into an instance segmentation, the semantic segmentation and boundary predictions were binarized over a threshold. Boundaries were then subtracted from the semantic segmentation before labeling connected components. The watershed algorithm, using the connected components as markers, was then applied in 2D or 3D in order to label the pixels or voxels in the boundary prediction.

Benchmark inference and evaluation

For the MitoNet and MitoNet-mini models, ortho-plane inference[46] was used on all benchmark volumes. During inference over each plane, median semantic segmentation probabilities were calculated from a queue of a few consecutive slices. A queue length of 3 was used for the C. elegans and fly brain, 5 for the HeLa and glycolytic muscle and 7 for the Lucchi++ and salivary gland. Queue lengths roughly track inversely with voxel size. Median probabilities were hardened at a confidence threshold of 0.3 and Panoptic-DeepLab postprocessing was performed. The resultant instance segmentations were passed through a connected components filter. After forward and backward instance matching with IoU and IoA threshold of 0.25, a size and bounding box extent filter were applied to eliminate likely false positives. Minimum sizes were 500 voxels for C. Elegans and fly brain; 800 for HeLa; 1,000 for salivary gland; 3,000 for glycolytic muscle; 5,000 for Lucchi++. Minimum bounding box extent was fixed at eight voxels. The same consensus segmentation algorithm as above was applied to ensemble the three segmentation stacks created by ortho-plane

inference. A clique IoU threshold of 0.75 and a vote threshold of 2 out of 3 was used for all benchmarks except the C. elegans and salivary gland volumes. For those a vote threshold of 1 out of 3 was used. This was in response to the relatively low IoU scores observed. While a vote threshold of 1 was used, each clique was required to have at least 2 detections; this enabled increases in IoU score without the introduction of many FPs. The consensus algorithm occasionally produced overlapping instances. Instances overlapping more than 100 voxels were merged; otherwise, the few overlapping voxels would be assigned to whichever instance was processed last.

For the Unet-BC model, ortho-plane inference was used on all benchmark volumes. First, semantic and boundary segmentation probabilities from xy, xz, and yz stacks were averaged together in 3D and binarized at thresholds of 0.3 and 0.2, respectively. Next, the boundary segmentation was subtracted from the semantic segmentation before labeling connected components. Finally, connected components larger than a minimum size were used as markers for the watershed algorithm. Minimum sizes for each volume were the same as those used for MitoNet postprocessing detailed above.

MitoNet finetuning

Finetuning was performed by manually selecting ground truth labeled patches of size 256^2 pixels from each of the six volume EM benchmarks. The number of patches used per volume were five for C. Elegans, fly brain, and HeLa cell (~2% of all voxels); eight for the glycolytic muscle (~0.5%); 16 for Lucchi++ (~0.8%); and 64 for salivary gland (~0.3%). The same hyperparameters used for MitoNet training were also used for finetuning with the exception that the max learning rate was set to 0.001 and the batch size was 16. All finetuned models were trained for 100 epochs.

OpenOrganelle volume inference and analysis

For the "Mouse kidney" dataset, two ROIs containing cells in the proximal tubule and distal tubule exclusively (some endothelial cells could not be avoided) were manually chosen from the 128nm overview to download the desired data at 16nm resolution. One ROI from the "Mouse liver" dataset was similarly selected. Inference used the best version of the MitoNet model and was run only on slices from the xy plane with a segmentation confidence threshold of 0.3 and median filter size of 5. A single compute node with a P100 GPU and 16 GB of memory was used. Image volumes and predicted mitochondrial instances were downsampled to 32nm pixel size for visualization and cleanup and a size threshold of 1,000 voxels was applied. For cleanup and quantification, 400 instances were randomly chosen such that they did not touch the boundary and each bounding box was at least 15 pixels long on a side. The empanada napari plugin was used to proofread each instance and merge, split, paint, and erase as required. FPs, mitochondria in endothelial cells, and instances on the boundary after merging were removed. Volume, surface area, flatness and elongation measurements for individual mitochondria were calculated using the LabelShapeStatistics filter implemented in ITK.[73] Branch lengths and cross-sectional diameters were calculated by first skeletonizing each mitochondrion and computing its distance transform. Lengths were simply the number of voxels in each branch, and the mean cross-sectional diameter was the average of distance transform values that overlapped with the skeleton. Branches shorter than 60 nm were pruned. Mitochondria contrast was calculated as the difference between maximum and minimum intensity voxels in each instance. Basolateral surfaces were approximated by fitting a triangular mesh to manually placed periodic fiducials using the ball pivoting algorithm[74] and the minimum distances between vertices of each mitochondrial mesh to the nearest basolateral surface were calculated.

QUANTIFICATION AND STATISTICAL ANALYSIS

Sample sizes and statistical tests used are reported in figure and table legends. The decision to use a parametric or non-parametric statistical test was made based on the observed normality of measured quantities. The threshold for statistical significance was 0.05.

ADDITIONAL RESOURCES

Links and additional explanations of all resources presented in this work are available on our project webpage https://volume-em.github.io/empanada.