MitoEM 2.0 - A Benchmark for Challenging 3D Mitochondria Instance Segmentation from EM Images

Peng Liu[1], Boyu Shen[1], Liyuan Liu[1], Qiong Wang[1], Shulin Zhang[3], Abhishek Bhardwaj[2], Ignacio Arganda-Carreras[4,5], Kedar Narayan[2], and Donglai Wei[1]

[1]Boston College
[2]National Cancer Institute
[3]Stanford University
[4]Donostia International Physics Center
[5]University of the Basque Country

Abstract

We present MitoEM 2.0, a curated data resource for training and evaluating three-dimensional (3D) mitochondria instance segmentation in volume electron microscopy. The collection assembles multiscale vEM datasets (FIB-SEM, SBF-SEM, ssSEM) spanning diverse tissues and species, with expert-verified instance labels emphasizing biologically difficult scenarios, including dense mitochondrial packing, hyperfused networks, and thin filamentous connections with ambiguous boundaries. All releases include native-resolution volumes and standardized processed versions, per-volume metadata (voxel size, modality, tissue, splits), and official train/validation/test partitions to enable reproducible benchmarking. Annotations follow a consistent protocol with quality checks and instance reindexing. Data are provided in NIfTI with nnU-Net–compatible layout, alongside machine-readable split files and checksums. Baseline scripts support common training pipelines and size-stratified evaluation. By consolidating challenging volumes and harmonized labels, MitoEM 2.0 facilitates robust model development and fair comparison across methods while supporting reuse in bioimage analysis, algorithm benchmarking, and teaching.

Background & Summary

Mitochondria are essential organelles that regulate cellular energy and metabolism. Their three-dimensional structure and organization are crucial for understanding cell function and disease. Volume electron microscopy (vEM) now enables high-resolution 3D imaging of mitochondrial ultrastructure, advancing the quantitative analysis of organelle architecture across cell types. However, automated analysis remains limited by current instance segmentation models, which struggle with challenging cases such as morphologically complex mitochondria, hyperfused networks, and tightly packed organelles with ambiguous boundaries. Although existing benchmarks [1, 2, 3] focus on simpler examples that inflate performance metrics, these difficult scenarios are where automated tools could have the greatest impact. In particular, although the previous MitoEM dataset [4] scaled 3D annotations by more than two orders of magnitude over previous efforts, it provided limited coverage of such difficult cases in hypermetabolic tissues where mitochondrial complexity is most pronounced. This gap continues to hinder the development of robust and generalizable models for biomedical discovery.

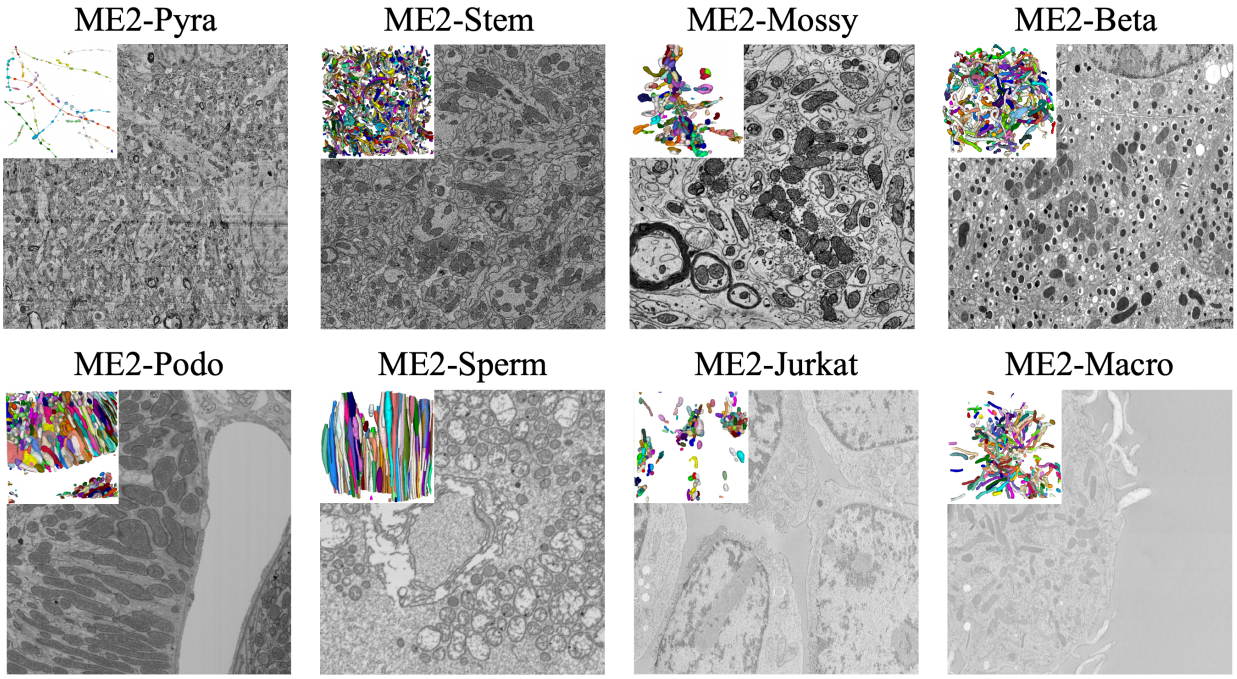

Figure 1: An overview of MitoEM 2.0.

To address these limitations, we introduce **MitoEM 2.0**, a curated benchmark targeting high-difficulty scenarios in 3D mitochondria segmentation (overview in Fig. 1) , with the following contributions:

- **Benchmark construction:** We select vEM tissues from known hypermetabolic tissue regions to form MitoEM 2.0, capturing complex morphologies such as hyperfused networks and ambiguous boundaries. We designed the benchmark datasets with paired training and test volumes for each region, along with standardized splits, metadata, and evaluation metrics.
- **Metrics for mitochondria instance hardness:** We developed two quantitative metrics to assess segmentation difficulty based on morphological complexity and spatial proximity. Analysis of existing datasets reveals a lack of coverage for challenging cases.
- **Reproducibility:** To facilitate reproduction, we open source the code for baseline models.

Methods

Benchmark Construction. The MitoEM 2.0 dataset catalog is organized around biological systems and cell types that pose known segmentation challenges arising from: (1) high mitochondrial density (crowding and small inter-mitochondrial spacing), (2) morphological complexity (branching, fenestrations, irregular surfaces), and (3) restrictive spatial context, including tight packing, confinement in narrow neurites or myofibrillar spaces, and close apposition to membranes/organelles. Our design prioritizes regions where mitochondria exhibit extreme morphological variability and where automated segmentation algorithms commonly fail, ensuring the benchmark captures the full spectrum of real-world segmentation difficulty.

Dataset Standardization: All volumes were acquired using volume electron microscopy (EM) modalities, including serial block-face scanning EM (SBF-SEM) and focused ion beam scanning EM (FIB-SEM), across multiple biological specimens. For near-isotropic FIB-SEM data, we downsampled the image volumes to 16nm with linear interpolation and antialising. For anisotropic ssSEM data, we preserved the original 8×8×30 nm. All datasets include metadata describing the species, tissue type, imaging modality, and voxel resolution.
Annotation Procedure: Mitochondria were annotated using a two-step semi-automated pipeline. First, initial segmentations were generated using a fine-tuned 3D U-Net[5] trained on a subset of the data. Next, expert annotators proofread the output using the VAST Lite and Neuroglancer tools, manually correcting under-segmentation, over-segmentation, and merge errors. All mitochondria with a minimum volume of 500 voxels were included. Each mitochondrion was assigned a unique instance label and verified across contiguous slices to ensure topological consistency. To ensure annotation quality, a subset of volumes was independently reviewed by multiple annotators. Discrepancies were resolved through consensus or exclusion. In total, over 8,000 mitochondria instances were labeled across 28 volumes.

Dataset Details. This section provides detailed descriptions of the eight constituent datasets in the MitoEM 2.0 benchmark, outlining their biological sources, imaging modalities, and annotation scales. The benchmark integrates volumetric electron microscopy (EM) datasets acquired using three modalities—FIB-SEM, ssSEM, and BF-SEM—spanning multiple organisms and tissues. These datasets capture substantial variability in mitochondrial density, morphology, and imaging resolution, forming a comprehensive resource for evaluating instance-level segmentation. Table 1 summarizes all datasets included in the benchmark, including their biological sources, imaging modalities, and annotation statistics.

- *ME2-Beta* (Müller *et al.*, 2020 [6]): Mouse pancreatic β-cells imaged by FIB-SEM (16×16×16 nm), comprising seven 874×669×979-voxel volumes with 1,935 annotated mitochondria. This dataset exhibits extremely high mitochondrial density and complex morphologies typical of insulin-secreting cells.

- *ME2-Jurkat* (Xu *et al.*, 2021 [7]): Human Jurkat T-cells imaged by FIB-SEM (16×16×16 nm), consisting of three 1024×1024×256-voxel volumes with 372 annotated mitochondria, representing mitochondrial organization in immortalized human lymphocytes used in immunological studies.

Name	Organism	Tissue	Cell Type	Modality	Voxels	Split	#Mitos	Source	EFI/DCI
ME2-Beta	Mouse	Pancreas	β-cell	FIB-SEM	3961M	3/1/3	1935	[6]†	1.02/2.68
ME2-Jurkat	Human	Cell line	Jurkat	FIB-SEM	805M	1/1/1	372	[7]*	1.00/1.62
ME2-Macro	Human	Cell line	Macrophages	FIB-SEM	805M	1/1/1	497	[7]*	1.01/1.38
ME2-Mossy	Mouse	Cerebellum	Mossy Fiber	ssSEM	3333M	3/1/3	1099	[8]	1.10/4.13
ME2-Pyra	Human	Cerebrum	Pyramidal	ssSEM	8388M	3/1/11	45	[4]	1.96/0.53
ME2-Podo	Mouse	Kidney	Podocyte	FIB-SEM	805M	1/1/1	1333	[9]*	1.00/2.43
ME2-Sperm	Drosophila	Gonad	Spermatocyte	FIB-SEM	1073M	3/1/4	775	[10]†	1.01/0.96
ME2-Stem	Mouse	Brainstem	Mix	BF-SEM	300M	1/1/1	2058	[11]	1.09/0.80

Table 1: Proposed MitoEM 2.0 benchmark with challenging mitochondria segmentation datasets based on biologically complex regions with dense or morphologically diverse mitochondrial populations. The *Split* column reports the approximate train/validation/test ratio in terms of annotated voxel volume. All FIB-SEM datasets have a resolution of 16×16×16 nm, whereas ssSEM and BF-SEM datasets have 8×8×30 nm resolution. Existing annotations were refined or converted from semantic masks (*), and new ones were created (†).

- *ME2-Macro* (Xu *et al.*, 2021 [7]): Human macrophage cell lines imaged by FIB-SEM (16×16×16 nm), comprising three 1024×1024×256-voxel volumes with 497 annotated mitochondria. The dataset captures interconnected mitochondrial networks that support phagocytosis and inflammatory responses.

- *ME2-Mossy* (Han *et al.*, 2024 [8]): Mouse cerebellar mossy-fiber boutons imaged by ssSEM (8×8×30 nm), including five 1024×1024×500-voxel volumes with 1,099 annotated mitochondria. It highlights densely packed presynaptic mitochondria supporting high-energy synaptic transmission.

- *ME2-Pyra* (Wei *et al.*, 2020 [4]): Human cortical pyramidal neurons imaged by ssSEM (8×8×30 nm), containing one 4096×4096×500-voxel volume with 45 annotated mitochondria. Despite their relatively small number, these mitochondria display high morphological diversity and intricate 3D organization within dendrites and somata.

- *ME2-Podo* (CellMap Consortium, 2024 [9]): Mouse kidney podocytes imaged by FIB-SEM (16×16×16 nm), including two 1024×1024×256-voxel volumes with 1,333 annotated mitochondria. The dataset depicts intricate mitochondrial networks within the specialized filtration structures of glomerular podocytes.

- *ME2-Sperm* (Kunduri *et al.*, 2022 [10]): *Drosophila* spermatocytes imaged by FIB-SEM (16×16×16 nm), comprising two 1024×1024×512-voxel volumes with 775 annotated mitochondria. It illustrates distinct mitochondrial arrangements and morphogenetic transitions during spermatogenesis.

- *ME2-Stem* (Jiang *et al.*, 2025 [11]): Mouse brainstem tissue imaged by BF-SEM (8×8×30 nm), comprising three 1000×1000×100-voxel volumes with 2,058 annotated mitochondria. This dataset features mixed neural and glial populations and moderate mitochondrial densities representative of brainstem tissue.

Each dataset entry is accompanied by standardized metadata describing its biological source (species, organ, cell type), imaging modality (FIB-SEM, ssSEM, BF-SEM), resolution, volume dimensions, and train/validation/test splits. The catalog spans multiple species (human, mouse, *Drosophila*) and organ systems, with voxel resolutions ranging from 8×8×30 to 16×16×16 nm and volume sizes from 1024×1024×256 to 4096×4096×500. This unified organization facilitates systematic benchmarking of segmentation models across distinct biological domains, enabling robust evaluation under diverse morphological and imaging conditions.

Metrics for Mitochondria Instance Hardness. To systematically assess the difficulty of mitochondria segmentation across diverse datasets, we introduce two complementary per-instance metrics: the *Dilation Collision Index (DCI)* and the *Erosion Fragility Index (EFI)*. Together these quantify whether an instance is likely to contact neighbors under small expansions (DCI) and whether its topology is fragile under small contractions (EFI).

- *Dilation Collision Index (DCI)*: For each mitochondrion instance M_i we apply a small morphological dilation with kernel B_{r_d} (radius r_d) and record the number of distinct other instances that the dilated region touches. In implementation, we use a $3 \times 3 \times 3$ structuring element (i.e., $r_d = 1$ voxel, full 26-neighborhood connectivity). Formally,

$$\mathrm{DCI}_i \;=\; \#\big\{\, j \neq i \mid \mathrm{Dilate}(M_i, B_{r_d}) \cap M_j \neq \varnothing \,\big\},$$

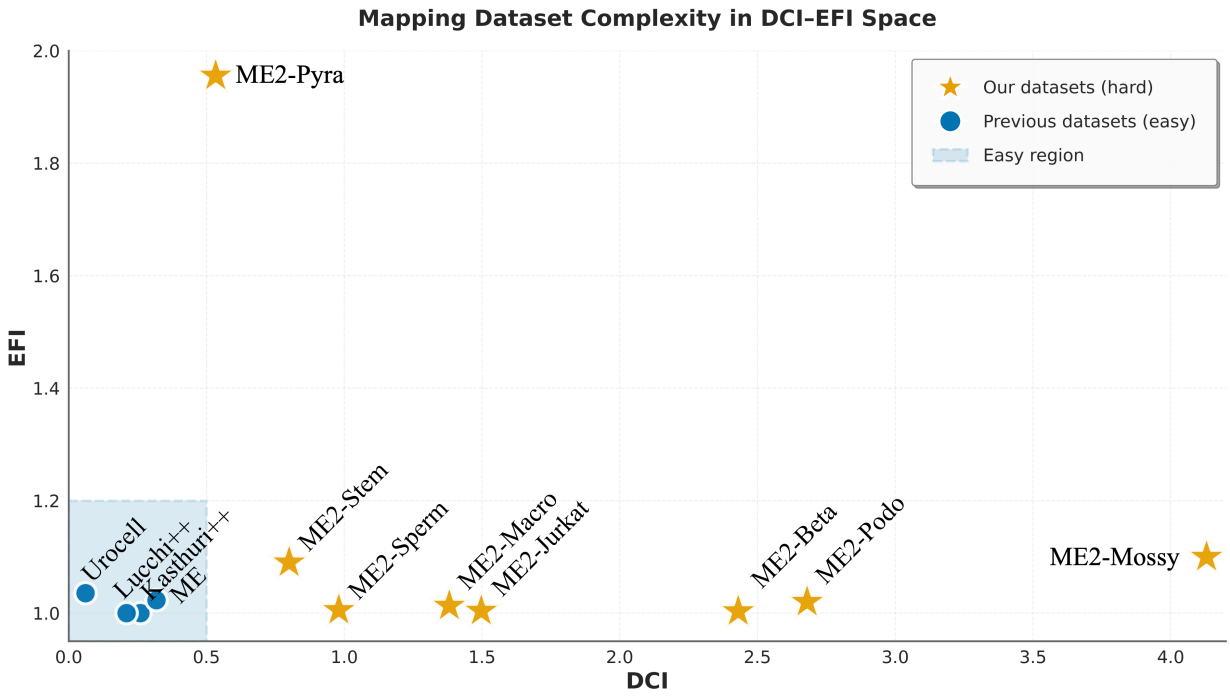

Figure 2: Instance segmentation hardness for the test split of MitoEM 2.0 and prior benchmarks.

i.e. the cardinality of the set of other instance labels overlapping the dilated mask. A larger DCI indicates that a small boundary expansion would cause contact with many neighbors, i.e., higher likelihood of false merges or boundary ambiguity. Practical computation should ensure unique-label counting (no double counts) and choose a dilation kernel that reflects expected boundary uncertainty.

- *Erosion Fragility Index (EFI)*: For each instance M_i we apply a mild morphological erosion with kernel B_{r_e} (radius r_e) and count the resulting number of connected components. Similarly, we use a 3×3×3 structuring element ($r_e = 1$ voxel, full 26-neighborhood).

$$\mathrm{EFI}_i \;=\; N_{\text{components}}\big(\mathrm{Erode}(M_i, B_{r_e})\big).$$

EFI captures topological fragility: values greater than 1 indicate that a slight inward perturbation tends to split the instance into multiple pieces (common for thin or highly branched mitochondria), implying a higher risk of fragmentation or undersegmentation during automatic processing.

By projecting mitochondria into the 2D space defined by $(\mathrm{DCI}, \mathrm{EFI})$ — we reveal clusters of instances with differing segmentation difficulty: high-DCI/low-EFI objects are densely packed but topologically robust, whereas low-DCI/high-EFI objects are isolated but topologically fragile, for example exhibiting constrictions, thin extensions, or mitochondrial nanotunnels. Many existing benchmarks concentrate in the low-DCI, low-EFI region, leaving a gap in coverage of cases prone to merge- and split-errors. These metrics therefore provide a principled basis to select and evaluate hard instances, and are integral to MitoEM 2.0dataset design and evaluation framework. A visual comparison of complexity across benchmarks is provided in Fig. 2.

Data Records

The repository contains per-dataset folders with a nnU-Net–compatible layout: `imagesTr/, labelsTr/, imagesTs/, labelsTs/`, plus `dataset.json` (name, modality, voxel spacing, channels), `splits.json` (official train/val/test assignment), `metadata.csv` (system, organ/tissue, cell type, resolution, source), `LICENSE`, and a top-level `README`. Images are NIfTI (`*.nii.gz`) at native and standardized spacings; labels are instance masks. Each case ID matches across image and label files (e.g., `CASE001_0000.nii.gz`, `CASE001.nii.gz`). For anisotropic ssSEM, we provide both native anisotropic and auxiliary isotropic exports; all variants are referenced in `dataset.json`. Code and exact preprocessing parameters used to generate standardized exports are archived with the release.

Technical Validation

Baseline method. We evaluate three baseline instance-segmentation methods representative of current practice for mitochondria in EM: (1) *MitoNet* [3], a Panoptic-DeepLab based generalist model for mitochondria; (2) *nnUNet-BC* [4] binary mask + contour channel, and (3) *micro-SAM* [12], a microscopy-adapted version of the Segment Anything Model. For each method we ran two distinct experimental regimes to separate out generalization from dataset-specific learning:

- *MitoNet.* A Panoptic-DeepLab based generalist for mitochondria (with Empanada/napari tooling for inference and proofreading). Instance masks are produced by center/offset predictions and subsequent grouping/post-processing (optionally with ortho-plane consensus for 3D). (1) Zero-shot. run the published MitoNet pretrained model on our test volumes without further adaptation, using the authors' recommended inference/post-processing settings. (2) Fine-tuning. train or fine-tune MitoNet on our training split (same input format and augmentations as in Methods) and evaluate with the identical inference pipeline.

- *nnUNet-BC.* A nnU-Net style encoder–decoder that predicts a binary foreground mask together with an auxiliary boundary/contour channel ("BC"). Instance masks are obtained by combining the foreground prediction with the contour channel via watershed/connected-component post-processing to split touching objects. We train the nnUNet-BC model on our training split (same preprocessing and patching as the main experiments), then evaluate on held-out test blocks. Training uses the standard Dice/BCE losses and the same augmentation pipeline described in methods.

- *micro-SAM (µSAM).* A microscopy-fine-tuned SAM variant that supports both interactive prompts and automatic instance generation (AMG/AIS). For automatic instance segmentation we run the model's AIS (Automatic Instance Segmentation) pipeline to produce instances. (1) Zero-shot. use the µSAM pretrained EM organelle weights (no additional training on our dataset) and generate instances using AIS with the default parameter set; this measures out-of-domain generalization and the utility of interactive/foundation-model approaches without dataset specific training. (2) Fine-tuning. fine-tune EM organelle µSAM model and AIS head on our training split following the same protocol as in methods, then run AIS inference and identical post-processing on the test data.

For all baselines we keep the input/output file layout and spatial metadata identical to the rest of the study (compressed NIfTI / nnUNet layout) so that preprocessing, resampling, and visualization are comparable. Inference outputs from every baseline pass the same post-processing

steps (small-object removal, connected-component relabeling, and optional 3D merging rules) prior to evaluation. Each baseline is reported under both the *zero-shot* and *from-scratch / fine-tune* settings to quantify (i) off-the-shelf generalization and (ii) achievable performance after dataset-specific training; evaluation metrics follow those defined in the main methods.

Evaluation

Following the recommendations and retrospective analysis in Franco-Barranco et al. [13], we adopt **matching-based accuracy** as the primary ranking metric for instance segmentation; this choice better accounts for false negatives and over-/under-segmentation in large 3D EM volumes. Predicted instances $P = \{p_i\}$ are matched one-to-one to ground-truth instances $G = \{g_j\}$ using the Hungarian algorithm on an IoU-based cost matrix with a minimum IoU threshold τ. After optimal assignment we count true positives (TP: matched pairs with IoU$\geq \tau$), false positives (FP: unmatched predictions), and false negatives (FN: unmatched ground truth). Accuracy is defined as

$$\text{Accuracy} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}.$$

By default we use $\tau = 0.5$ for the primary matching, but results for alternative τ values are reported where relevant. The precise cost construction (IoU or $1 - \text{IoU}$), tie-breaking, and assignment heuristics are implemented in the released evaluation code.

Benchmark on each dataset

For each benchmark volume, we evaluate three representative baselines—**nnUNet-BC**, **MitoNet**, and **Micro-SAM**—under two regimes: *zero-shot (ZS)* inference and *fine-tuned (FT)* training. Quantitative results are summarized in Table 2. For nnUNet-BC, which does not rely on pre-trained weights, the corresponding regime is denoted as *Scratch*, indicating training from random initialization on each dataset's designated training split. All models are evaluated using identical preprocessing, postprocessing, and evaluation metrics.

Evaluation protocol.

1. **Inference mode.** In the *zero-shot* setting, pretrained models (MitoNet, Micro-SAM) are directly applied to held-out benchmark volumes without adaptation. In the *from-scratch/fine-tuned* setting, models are either trained from scratch (nnUNet-BC) or lightly fine-tuned (MitoNet, Micro-SAM) on each dataset training data and then evaluated on the test split.

2. **Ortho-plane or volumetric inference.** For models supporting orthogonal-plane inference (e.g., MitoNet), we adopt the same slice-queue and median-filtering strategy from Conrad & Narayan [3], adapting queue lengths (3–7) to voxel resolution.

3. **Metric definition.** We report two complementary accuracies: the first number (**left**) represents *instance-level accuracy* (object-wise matching), and the second number (**right**) represents *semantic-level accuracy* (voxel-wise overlap). Both are computed using Hungarian one-to-one matching with IoU threshold $\tau = 0.50$.

Table 2: Per-dataset `Accuracy@50`. The first half reports **semantic-level** accuracy and the second half reports **instance-level** accuracy (Hungarian matching, IoU threshold $\tau = 0.50$). Bold indicates best performance per row for each metric.

Dataset	Semantic-level Accuracy (↑)					Instance-level Accuracy (↑)				
	nnUNet-BC	MitoNet		Micro-SAM		nnUNet-BC	MitoNet		Micro-SAM	
	Scratch	ZS	FT	ZS	FT	Scratch	ZS	FT	ZS	FT
ME2-Beta	**92.8**	80.2	88.0	64.3	89.5	16.9	35.0	**54.4**	8.5	34.3
ME2-Jurkat	**84.9**	78.1	76.1	77.6	77.4	**59.6**	44.0	52.4	30.4	34.6
ME2-Macro	**89.5**	74.4	87.0	78.5	87.0	30.5	30.7	**43.3**	16.3	40.2
ME2-Mossy	83.4	40.9	68.9	89.6	**89.9**	6.6	6.1	27.0	34.4	**36.9**
ME2-Podo	90.4	**91.6**	**91.6**	87.4	91.0	27.3	35.4	**58.3**	25.7	39.0
ME2-Pyra	**91.8**	90.0	89.0	89.2	89.4	16.9	17.5	**23.2**	16.3	18.0
ME2-Sperm	91.6	85.9	92.0	86.9	**92.7**	60.8	16.8	**63.7**	15.6	34.6
ME2-Stem	71.7	67.0	72.5	71.9	**79.8**	**37.3**	31.8	35.5	25.7	31.9
Average	87.0	76.0	83.1	80.7	**87.1**	32.0	27.2	**44.7**	21.6	33.7

Overall comparison and insights. Across all eight benchmark volumes, **MitoNet (FT)** achieves the highest overall performance (instance/semantic = 44.7/83.1), indicating that even minimal fine-tuning can substantially improve both object detection and pixel-level precision. Its zero-shot variant (27.2/76.0) demonstrates moderate generalization but clear sensitivity to domain shift, underscoring the importance of limited adaptation.

nnUNet-BC (Scratch) achieves the second best semantic accuracy (87.0) among all methods, showing that fully supervised training on domain-specific data enables precise boundary delineation. However, its lower instance-level accuracy (32.0) suggests more frequent over- or under-segmentation of adjacent mitochondria, consistent with its lack of pretraining priors.

Micro-SAM (FT) closes much of the gap to nnUNet-BC in both metrics (33.7/87.1), though its zero-shot variant (21.6/80.7) exhibits weaker instance detection—likely reflecting the domain mismatch between its natural-image pretraining and EM data.

Overall, these results highlight a clear pattern: (1) domain-pretrained models like MitoNet exhibit superior instance recognition and transferability; (2) fully supervised nnUNet-BC and finetuned Micro-SAM excels at fine-grained voxel delineation; and (3) general foundation models (Micro-SAM) require moderate fine-tuning to bridge cross-domain gaps. While many methods achieve strong semantic-level accuracy, their instance-level accuracy on the MitoEM 2.0 (ME2) datasets remains substantially lower, revealing frequent split/merge failures; this gap underscores the need to assess both object completeness and boundary precision rather than voxel overlap alone.

Data Availability

The curated dataset is available at `https://huggingface.co/datasets/pytc/MitoEM2.0`.

Code Availability

All evaluation code (instance matching, metric computation, thresholds and heuristics) will be released with the dataset so that reported numbers are fully reproducible and directly comparable to prior benchmarks. The retrospective re-ranking and metric discussions in [13] motivate providing

the exact implementation alongside the data. All code is released under the MIT License and is compatible with Python 3.9 and PyTorch 1.12+. Installation instructions and reproducibility notes are provided in the project README file.

References

[1] A. Lucchi, P. Márquez-Neila, C. Becker, Y. Li, K. Smith, G. Knott, and P. Fua, "Learning structured models for segmentation of 2-D and 3-D imagery," *IEEE Transactions on Medical Imaging*, vol. 34, no. 5, pp. 1096–1110, 2014.

[2] V. Casser, K. Kang, H. Pfister, and D. Haehn, "Fast mitochondria detection for connectomics," in *Medical Imaging with Deep Learning*, 2020.

[3] R. Conrad and K. Narayan, "Instance segmentation of mitochondria in electron microscopy images with a generalist deep learning model trained on a diverse dataset," *Cell Systems*, vol. 14, no. 1, pp. 58–71, 2023.

[4] D. Wei, Z. Lin, D. Franco-Barranco, N. Wendt, X. Liu, W. Yin, X. Huang, A. Gupta, W.-D. Jang, X. Wang *et al.*, "MitoEM dataset: Large-scale 3D mitochondria instance segmentation from EM images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2020, pp. 66–76.

[5] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

[6] A. Müller, D. Schmidt, C. S. Xu, S. Pang, J. V. D'Costa, S. Kretschmar, C. Münster, T. Kurth, F. Jug, M. Weigert *et al.*, "3d fib-sem reconstruction of microtubule–organelle interaction in whole primary mouse β cells," *Journal of Cell Biology*, vol. 220, no. 2, p. e202010039, 2020.

[7] C. S. Xu, S. Pang, G. Shtengel, A. Müller, A. T. Ritter, H. K. Hoffman, S.-y. Takemura, Z. Lu, H. A. Pasolli, N. Iyer *et al.*, "An open-access volume electron microscopy atlas of whole cells and tissues," *Nature*, vol. 599, no. 7883, pp. 147–151, 2021.

[8] X. Han, X. Lu, P. H. Li, S. Wang, R. Schalek, Y. Meirovitch, Z. Lin, J. Adhinarta, K. D. Murray, L. M. MacNiven *et al.*, "Multiplexed volumetric clem enabled by scfvs provides insights into the cytology of cerebellar cortex," *Nature communications*, vol. 15, no. 1, p. 6648, 2024.

[9] C. P. Team, D. Ackerman, M. B. Ahrens, Y. Aso, E. Avetissian, D. Bennett, C. K. E. Bleck, J. Bogovic, M. Bryant, D. Crooks, D. Feliciano, J. Funke, L. Heinrich, H. Hess, M. Innerberger, N. Iyer, M. Kittisopikul, W. Korff, W.-P. Li, W. M. Linehan, Z. Lu, S. Pang, W. Park, K. Pedram, A. Petruncio, A. Post, S. Preibisch, J. Price, W. Qiu, D. Ramirez, J. Rhoades, V. M. S. Ruetten, S. Saalfeld, E. T. Trautman, R. Vorimo, A. Weigel, C. S. Xu, G. Yu, Z. Yu, M. Zouinkhi, and Y. Zubov, "CellMap Segmentation Challenge," 12 2024. [Online]. Available: https://janelia.figshare.com/articles/online_resource/CellMap_Segmentation_Challenge/28034561

[10] G. Kunduri, S.-H. Le, V. Baena, N. Vijaykrishna, A. Harned, K. Nagashima, D. Blankenberg, I. Yoshihiro, K. Narayan, T. Bamba *et al.*, "Delivery of ceramide phosphoethanolamine lipids to the cleavage furrow through the endocytic pathway is essential for male meiotic cytokinesis," *PLoS Biology*, vol. 20, no. 9, p. e3001599, 2022.

[11] Y. Jiang, H. Wang, K. M. Boergens, N. Rzepka, F. Wang, and Y. Hua, "Efficient cell-wide mapping of mitochondria in electron microscopic volumes using webknossos," *Cell Reports Methods*, vol. 5, no. 2, 2025.

[12] A. Archit, L. Freckmann, S. Nair, N. Khalid, P. Hilt, V. Rajashekar, M. Freitag, C. Teuber, M. Spitzner, C. Tapia Contreras *et al.*, "Segment anything for microscopy," *Nature Methods*, vol. 22, no. 3, pp. 579–591, 2025.

[13] D. Franco-Barranco, Z. Lin, W.-D. Jang, X. Wang, Q. Shen, W. Yin, Y. Fan, M. Li, C. Chen, Z. Xiong *et al.*, "Current progress and challenges in large-scale 3d mitochondria instance segmentation," *IEEE transactions on medical imaging*, vol. 42, no. 12, pp. 3956–3971, 2023.

Author Contributions

D.W. conceived the project and supervised dataset design and curation. P.L., B.S., L.L, Q.W., S.Z and A.B performed the data preprocessing and annotation. P.L. developed the baseline segmentation models and benchmarking pipeline. B.S. and L.L. contributed to annotation tool integration and quality control. P.L. and D.W. conducted technical validation and evaluation. D.W., P.L., K. N., and I.A-C. wrote the manuscript with input from all authors. All authors reviewed and approved the final version of the manuscript.

Competing Interests

The authors declare no competing interests.

Acknowledgments

Acknowledgments

This work was supported by the NSF-IIS-2239688. It was also partially supported by grant PID2024-157485NB-I00, funded by the Ministerio de Ciencia, Innovación y Universidades (MICIU/AEI /10.13039/501100011033/) and "ERDF – A way of making Europe" (to I.A-C.).

Funding

Ethics statement