OmniEM: Unifying the EM Multiverse through a Large-scale Foundation Model

Liuyuan He[1,*], Ruohua Shi[2,*], Wenyao Wang[1,*,+], and Lei Ma[1,3,+]

[1]National Biomedical Imaging Center, College of Future Technology, Peking University, Beijing, 100871, China
[2]National Engineering Research Center of Visual Technology, School of Computer Science, Peking University, Beijing, 100871, China
[3]Beijing Academy of Artificial Intelligence, Beijing, 100084, China
[*]these authors contributed equally to this work
[+]wangwenyao@pku.edu.cn
[+]lei.ma@pku.edu.cn

ABSTRACT

Room-temperature electron microscopy (EM) analysis is rapidly evolving but faces growing challenges due to massive data heterogeneity and fragmented methodologies, forming a significant bottleneck in advancing EM-based discovery. To overcome this, we present OmniEM, the first unified and extensible foundation model framework that consolidates a wide spectrum of EM tasks under one platform. At its core is EM-DINO, the first and most powerful EM-specific vision foundation model trained via large-scale self-supervised learning (SSL) on EM-5M, the largest and most comprehensive EM dataset to date, containing 5 million standardized images spanning diverse species, tissues, preparation protocols, imaging modalities, and resolutions ranging from 0.5 nm to 70 nm. Notably, this dataset was curated through a novel model-driven filtering protocol. EM-DINO demonstrates strong EM image understanding ability in semantic classification, and OmniEM extends the capacity of EM-DINO through a lightweight yet powerful U-shaped architecture, enabling impressive performance in robust mitochondrial segmentation, multi class subcellular segmentation and more. By unifying data, model, and software into a cohesive foundation, OmniEM marks a critical step toward general-purpose EM image understanding. By consolidating multiple tasks into a single framework, OmniEM streamlines EM analysis, representing a significant step toward more efficient and reusable EM workflows that accelerate scientific discovery. The data and code for this paper will be released soon.

Main

Electron microscopy (EM) has revolutionized biological discovery by enabling ultra-high-resolution, stereoscopic visualization of subcellular and tissue-level structures.[1,2]. It is the definitive technology for decoding biological complexity at the nanoscale, powering breakthroughs in neuroscience, cell biology, and beyond. From the complete reconstruction of neural circuits in *Caenorhabditis elegans*[3–5], *Drosophila*[6–9], zebrafish[10–12], and mice[13–17], to recent mapping cortical connectivity in non-human primates and human[18,19], EM has fundamentally reshaped our understanding of the brain. Beyond neuroscience, EM has become indispensable across disciplines, driving transformative insights into cellular architecture and organelle dynamics[20,21] and numerous other research fields[22]. As EM enters the era of high-throughput and 3D imaging, the demand for automated, intelligent, and scalable analysis pipelines has never been more urgent[23]. Yet, despite rapid technological advances in hardware, EM data analysis remains a bottleneck which calls for a paradigm shift in both methodology and infrastructure[24].

The explosive growth of EM data caused by advances in imaging technologies has pushed traditional analysis pipelines to their breaking point[25]. While deep learning (DL) methods have made notable progress[26], particularly in neuron segmentation for connectomics. Cutting-edge DL models have emerged for restoration,[27–29], registration[30,31], and especially segmentation[32–36], but these models are almost exclusively optimized for narrowly defined datasets and tasks, failing to generalize across the diverse biological and technical conditions that define real-world EM research. In practice, EM studies involve a vast spectrum of species, tissue types, fixation protocols, staining variations, and imaging modalities[37]. This exacerbates discrepancies between datasets and results in catastrophic performance degradation of trained models outside their training regime. The issue is compounded by the fact that EM workflows are inherently multi-step and modular, requiring distinct models for each stage, such as restoration, alignment, segmentation, etc.[38,39]. This make the current situation that precious time and resources of researchers are consumed not by scientific discovery, but by the tedious work of adapting and stitching together incompatible tools and models[40,41]. Worse still, some non-computational domain researchers are locked out from leveraging these powerful tools, as assembling a full-stack EM analysis pipeline demands significant engineering expertise. At this point, the current mess of tools and systems has become a roadblock hat keeps science from moving forward at the speed it should[42].

Multiple attempts have been made to simplify the reuse of deep models in biological imaging. Tools such as DeepMIB[43], DeepImageJ[44], and UNI-EM[40] ower the entry barrier by offering user-friendly interfaces, while modular frameworks like PyTorch Connectomics[45] and DACAPO[46] promote flexible pipeline construction. Yet these tools remain fundamentally constrained: users are still required to annotate domain-specific datasets and retrain models, consuming vast amounts of time and computational resources for each task. More recent efforts have turned toward generalization. For example, MitoNet[35] offers zero-shot segmentation for mitochondria, and EM-diffuse[29] introduces a diffusion-based suite for denoising and reconstruction. Despite their innovations, such models fail to break out of narrow task domains, lacking true general-purpose capability.

Meanwhile, vision foundation models (VFMs) have sparked a revolution in medical imaging[47], which have been successfully applied in radiology[48,49], magnetic resonance imaging[50], ophthalmology[51], pathology[52–55] and fluorescence microscopy[56], yet their application in EM analysis has not been systematically explored. Even though several fine-tuned Segment Anything Models(SAM)[57] have been proposed, such as MedSAM for medical images[58], microSAM for microscopy[59], ShapeMamba-EM for EM[60], DendriteSAM for dendrite[61], and TriSAM for cortical blood vessel[62]. Their dependency on user-provided prompts limits their use to auxiliary annotation rather than enabling fully automated analysis[63]. Moreover, the supervised training paradigm of SAM is specifically designed for spatial understanding, restricting its adaptability to other types of tasks[64]. As a result, the EM field still lacks a scalable, fully automated, and general-purpose EM foundation model capable of performing diverse analytical tasks across heterogeneous datasets, without retraining or prompt engineering. This absence is not merely a technical gap; it is a systemic failure to bring EM into the next era of intelligent, reusable, and foundation-model-driven analysis.

To this end, we propose a self-supervised learning (SSL) framework grounded in two core observations. First, the EM community has accumulated vast quantities of open-access raw image data[65], which, however, is not always accompanied by corresponding annotations. Second, SSL has consistently demonstrated its power in extracting transferable representations across diverse domains in computer vision[66], and the biomedical imaging[67], including recent progress in EM[35,68]. Yet, these strengths remain underexploited for building general-purpose EM models. To unlock this potential, we curated EM-5M, the largest and most diverse EM dataset to date, drawing from public repositories (Fig. 1 left and Extended Data Fig. 1). TEM-5M spans a wide spectrum of sample preparations, imaging modalities, resolutions, species, and tissue types, providing the scale and heterogeneity required for robust, transferable representation learning. Based on EM-5M, we pretrained EM-DINO, an EM-specific vision foundation model[69] through Self-distillation with no labels (DINOv2)[70], which could effectively capture EM-specific features (Fig. 1 middle). On top of EM-DINO, we introduced OmniEM, a unified, extensible framework for multi-task EM analysis (Fig. 1 right). Powered by EM-DINO as its encoder, OmniEM leverages a U-shaped decoder architecture to effectively integrate and propagate multi-scale features, enabling it to serve as a versatile, general-purpose solver across downstream tasks.

We perform extensive evaluations to validate our approach. First, EM-DINO exhibits superior semantic understanding of EM images, outperforming baseline DINOv2 across both in-domain and out-of-domain benchmarks (Fig. 2). Most critically, the learned representations from EM-DINO can be effectively transferred to multi-granularity downstream image analysis tasks through the OmniEM framework. Compared to previous leading methods, OmniEM achieves superior performance in semantic segmentation (Fig. 3). In summary, we present the first multi-task foundation model for biological EM domain. We envision that OmniEM will significantly enhance the efficiency of real-world EM data analysis and promote broader adoption of EM techniques within the biological research community.

Results

Schematic Overview of the Development Pipeline

The quality and content of the pretraining dataset critically determine a model's capacity for generalization. In the context of EM images, heterogeneity can be broadly categorized into two types: low-level heterogeneity, such as variations in brightness, contrast, and noise arising from differences in sample preparation, imaging devices, and acquisition conditions; and high-level heterogeneity, which involves semantic or geometric differences due to factors like species, organ types, and resolution. To enable the model to learn robust and generalizable EM representations, it is essential to construct a comprehensive dataset that captures a wide range of combinations across these factors. In this study, we first collected diverse raw datasets (approximately 500TB) from major public data-sharing platforms (EMPAIR, Openorganelle, BossDB, and WebKnossos), along with additional external sources. We then invested extensive efforts to reduce redundancy and ensure coverage of diverse biological and imaging conditions (detailed curation methods are described in the Methods section and Extended Data Fig. 1). As a result, we compiled a dataset of 5.5 million images, EM-5M, covering two primary sample preparation methods, five major electron microscopy modalities, 13 animal species, and 16 organ sources, with resolutions ranging from 0.5 nm to 70 nm (Fig. 1 left).

Using EM-5M, we pretrained the EM-specific foundation model, EM-DINO, starting from the pretrained DINOv2 checkpoint[70], in order to leverage DINOv2's superior ability to generate high-quality visual features at both the image and pixel levels (Fig. 1 middle). To adapt the model to the EM domain, we replaced the original DINOv2 data augmentation strategies

with domain-specific augmentations commonly used in EM image analysis[39,45] (details are described in the Methods section). To further support multiple downstream tasks, we integrated EM-DINO's multiscale feature representations into OmniEM, a U-shaped architecture designed to capture hierarchical EM features (Fig. 1 right). OmniEM was subsequently fine-tuned for specific applications across various EM analysis tasks. For practical use, we have packaged OmniEM's capabilities into a plugin for Napari, the most widely used and advanced biomedical image processing platform. In addition, to facilitate reproducibility and encourage broader community contributions, we have publicly released the source code for both the pretrained EM-DINO foundation model and the fine-tuned OmniEM model.

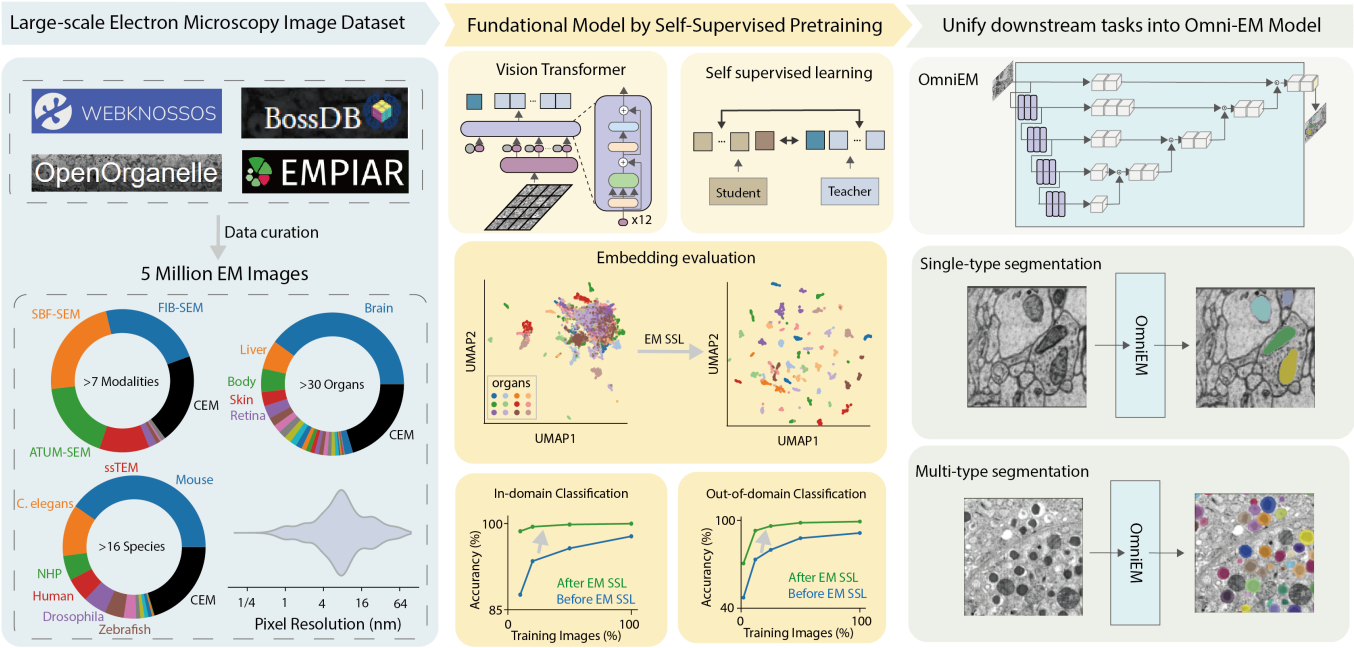


Fig. 1. Overview of EM-5M, EM-DINO, and OmniEM. Left. Data from Webknossos, BossDB, OpenOrganelle, EMPIAR and other databases are collected and curated to EM-5M, which contains 5 million standardized images spaning major imaging devices, tissue types, animal species and pixel resolutions. **Middle.** Illustration of the DINO SSL process and the improved classification ability of the model on both in-domain and out-of-domain EM data after training. **Right.** From top to bottom: the unified framework of OmniEM, different types of downstream tasks, and the Napari plugin interface.

EM-DINO captures contextual correlations between EM images

We hypothesize that EM-DINO can extract meaningful representations from EM images after being pretrained on the EM-5M dataset. To evaluate this, we conducted image classification tasks by appending a linear classifier head, and compared EM-DINO against a baseline DINOv2 model[70] pretrained solely on natural images. These classification tasks required models to distinguish between visually similar images originating from different organs.

For the in-domain evaluation, we sampled images from 11 organ types within the EM-5M dataset. Across various training data scales, EM-DINO consistently outperformed DINOv2, achieving notably high accuracy (>98%) even with as few as 200 labeled samples (Fig. 2 a). To assess zero-shot classification capability, we visualized the embedding spaces of both models using Uniform Manifold Approximation and Projection (UMAP)[71] without any finetuning (Fig. 2 b). The embeddings from EM-DINO formed well-separated clusters corresponding to distinct organs, whereas those from DINOv2 were more diffusely distributed—potentially explaining EM-DINO's superior classification performance after supervised finetuning.

To ensure generalization beyond the training domain, we further increased classification difficulty using out-of-domain images from 12 organ types sourced from Histology Guide[1], which have been excluded from EM-5M. These images had a higher physical resolution (<2 nm), posing challenges even for trained human annotators. Under these conditions, EM-DINO initially exhibited a drop in performance but rapidly recovered to exceed >95% accuracy (Fig. 2 c). In contrast, DINOv2 started from sharp accuracy drop (<50%) and failed to surpass 90% accuracy throughout. The corresponding zero-shot UMAP visualization (Fig. 2 d) exhibits more pronounced observation, in which DINOv2 embeddings lacked discernible structures but EM-DINO embeddings formed distinct sub-organ clusters. A deeper analysis revealed that EM-DINO tended to group images

[1] https://histologyguide.com/EM-atlas/EM-atlas.html

with similar fine-scale textures (Extended Data Fig. 2), indicating its ability to capture high-level semantic features relevant to EM image interpretation.

In conclusion, these findings demonstrate the effectiveness of domain-specific SSL in learning rich and generalizable representations from EM images. EM-DINO not only excels in familiar tasks but also generalizes robustly to novel datasets, supporting its potential as a foundational model for a wide range of downstream EM imaging applications.

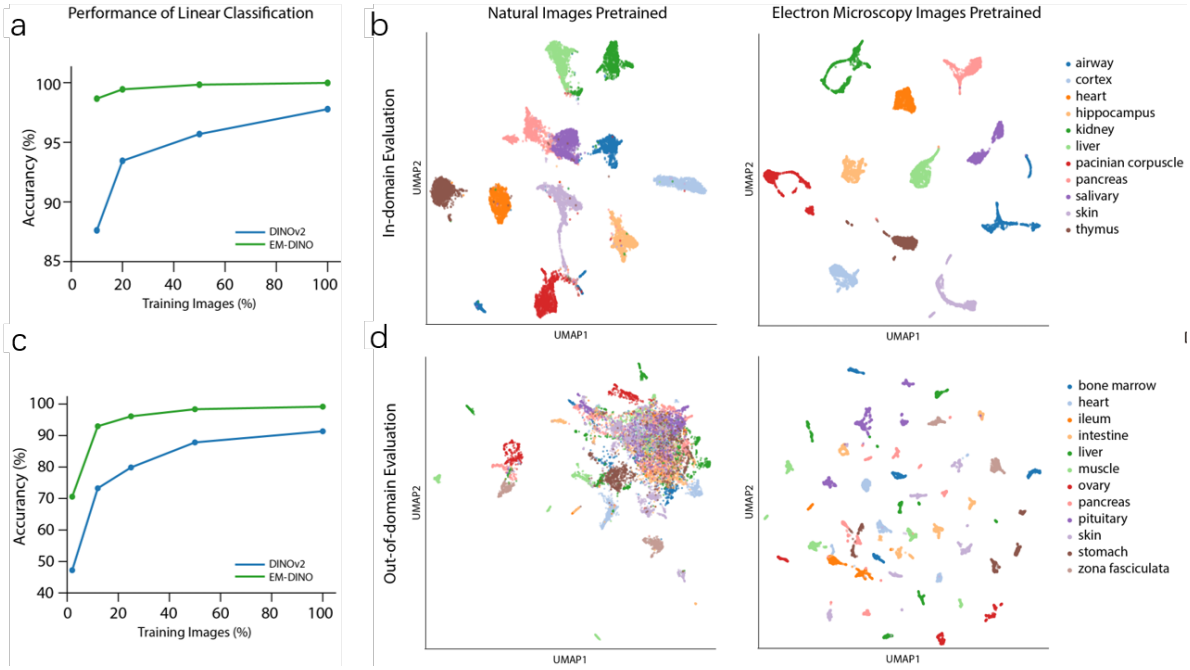


Fig. 2. Evolution of embeddings on organ-classification tasks. The EM-DINO model (pretrained with additional EM images) demonstrates superior EM image representation capability compared to DINOv2 (pretrained with only natural images) through SSL pretraining.**a** In-domain classification accuracy of both models relative to the amount of training data. **b** Zero-shot UMAP visualization of image embedding spaces across 11 organs on the in-domain classification task. **c** Out-of-domain classification accuracy of both models relative to the amount of training data. **d** Zero-shot UMAP visualization of image embedding spaces across 12 organs on the out-of-domain classification task.

Precious image segmentation

Image segmentation is a fundamental task in EM images processing, as it enables downstream analysis and deeper insights into biological systems. Among the various subcellular structures, mitochondria are frequently mentioned due to their distinctive morphology and their essential role in cellular metabolism and signaling. However, the appearance of mitochondria varies significantly across tissue types, which poses a substantial challenge for cross-dataset generalization[72]. MitoNet, a recently proposed general mitochondrial segmentation model, employs a ResNet backbone pretrained on CEM-1.5M and is fine-tuned with CEM-MitoLab, a curated dataset containing 21,860 manually annotated EM images[35]. In this study, we leveraged both the CEM-MitoLab dataset and the powerful representation capacity of EM-DINO to develop a mitochondrial segmentation branch within the OmniEM framework. Compared with MitoNet and a baseline DINOv2 model (not pretrained on EM data), our approach demonstrated superior zero-shot performance across seven benchmark datasets (Fig. 3 a and b). Additional improvements were observed with minimal fine-tuning (Fig. 3 c and d). Notably, while DINOv2 was not specifically pretrained on EM images, it achieved performance comparable to MitoNet on average, underscoring the potential of transformer for this task. After domain-specific pretraining on EM-5M, EM-DINO significantly outperformed both MitoNet and DINOv2, illustrating the value of a large, well-curated pretraining dataset in enhancing downstream performance. This strategy employed for mitochondrial segmentation can be readily extended to other organelles, provided that high-quality annotated datasets are available.

Compared to single-organelle segmentation, multi-class organelle segmentation poses significantly greater challenges due to the increased complexity of distinguishing multiple subcellular structures. Nevertheless, analyzing diverse organelles offers a more holistic understanding of cellular states and functions. With the rising adoption of EM techniques in cellular research, the demand for accurate multi-class segmentation continues to grow. To address this, we introduced a dedicated branch within

OmniEM for multi-class segmentation and evaluated its performance on the widely used Betaseg dataset[73]. Without employing any additional sophisticated strategies, OmniEM achieved consistent performance improvements across all organelle classes and in overall average accuracy (Fig. 3 e and f), outperforming prior leading methods. While multi-class segmentation in EM remains an emerging area, however we are seeing the expansion of multi-class segmentation datasets, such as the whole-cell segmentation challenge organized by CellMap(https://cellmapchallenge.janelia.org/). We anticipate that OmniEM's capabilities will scale with these datasets to support broader generalization and finer-grained semantic segmentation in the future.

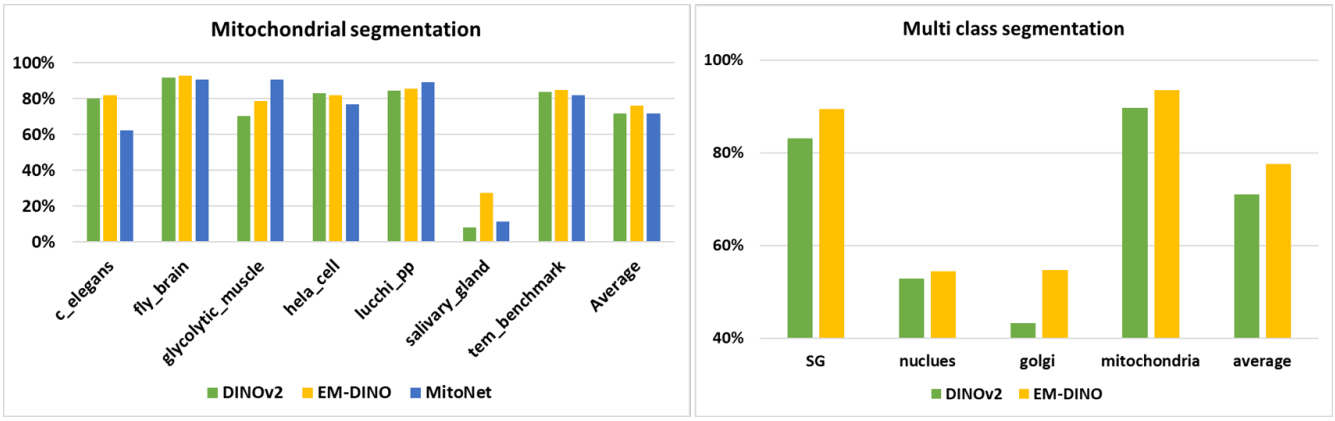


Fig. 3. Segmentation performance evaluation. a Zero-shot segmentation performance of EM-DINO compared to other methods on benchmark mitochondrial datasets. **b** Segmentation performance of EM-DINO versus DINOv2 on the BetaSeg datasets.

Discussion

In this study, we demonstrate the feasibility of leveraging large-scale, domain-specific SSL to construct a vision foundational model capable of unifying currently fragmented EM processing tasks which traditionally require long and complex pipelines. Such a unified framework, OmniEM, addresses a critical need in the EM community for simplified, efficient, and scalable image analysis tools, and aligns with the broader vision of AI for Science.

To support this goal, we curated and organized EM-5M, the largest and most comprehensive self-supervised training dataset currently available in the EM field. Building on this foundation, we developed and publicly released EM-DINO, which, to the best of our knowledge, represents the first self-supervised vision foundation model tailored specifically for EM imaging. Most importantly, we introduced OmniEM, a U-shaped encoder–decoder framework that effectively integrates the multi-scale features extracted by EM-DINO. In this way, OmniEM provides a unified solution for multiple EM image processing tasks, including denoising, super-resolution, and various kinds of segmentation.

Although achieving state-of-the-art performance was not the primary objective of this study, OmniEM nonetheless delivered compelling results across a variety of tasks. Notably, these outcomes were achieved without relying on any complex task-specific heuristics or architectures—highlighting the robustness and generalization ability of the proposed foundation model.

Major contribution and future direction

Our work makes three important contributions to EM analysis:

- The most prominent contribution of this work is EM-5M, the largest and most comprehensive dataset for EM to date, which was made possible by building upon the foundational efforts of the CEM-1.5M team and the broader open-source EM community. Our experiments confirm that the scaling law remains valid in the EM domain for that larger datasets continue to improve model performance. However, a critical insight should be noted is that data heterogeneity plays a more critical role than scale alone. Effective pretraining demands rigorous data curation to eliminate irrelevant and redundant content. To address this, we propose a model-driven data curation protocol, inspired by recent advances in DINOv2 and SAM. This iterative process involves pretraining on an initial dataset, using the model's feedback to identify redundancy, and progressively refining the dataset. While effective, this approach is computationally intensive. Future efforts should focus on developing efficient and automated filtering techniques to accelerate the construction of high-quality EM datasets.

- Another key contribution of this work is the EM-DINO model. OmniEM represents just one specific implementation built upon EM-DINO. We sincerely encourage the community to explore broader applications of EM-DINO by adapting it to their own tasks and workflows. Looking ahead, we plan to further extend EM-DINO through multi-modal learning, incorporating textual information to enhance clinical and pathological EM image interpretation. In parallel, we will explore the integration of diffusion models to generate high-quality synthetic EM data, which can, in turn, serve as valuable training resources to further improve model performance.

- The most immediate and practical outcome of this work is the OmniEM framework, which offers a suite of image processing capabilities and has been packaged as a Napari plugin for ease of use. Users can directly apply the provided task-specific checkpoints or follow our training protocols to finetune their customized models. For users with advanced requirements and development expertise, it is important to note that OmniEM is designed with extensibility, in which new task branches can be added seamlessly. Beyond using existing modules, users are encouraged to extend the framework by implementing their own task-specific branches. Our team will continue to expand OmniEM by incorporating additional task types, progressively evolving it into a more powerful and flexible general-purpose EM image processing solution.

Technology Considerations

In selecting between 2D and 3D format for pretraining, we ultimately opted for a 2D-based approach. This decision was guided by the availability of mature model architectures and well-established training frameworks in the 2D domain, which facilitate efficient learning from heterogeneous EM image sources. Although 3D pretraining strategies have been proposed, the high computational overhead and significant redundancy between adjacent slices outweigh the benefits of added axial context. Moreover, 2D models can be naturally extended to 3D applications, making them the more practical choice for developing a foundational model.

For the visual foundation model, we selected the DINOv2 checkpoint. Despite the widespread attention given to SAM for its outstanding segmentation performance, our aim was to build a versatile, multi-task model. DINOv2 demonstrated superior capabilities in general-purpose feature extraction, making it more suitable to support a broad range of downstream image analysis tasks.

Regarding the architecture for downstream tasks, we avoided using multiple task-specific heads, as seen in some previous biomedical foundation models. Instead, we adopted a U-shaped architecture, which has been commonly used in biomedical

image processing for its effectiveness in extracting and integrating multi-scale features. This choice not only enhanced task performance but also provided a unified architectural backbone across various "image-to-image" tasks.

Limitation and optimization directions

Our paper provides a reference for for the development of a general-purpose EM model, demonstrating competitive performance across a variety of tasks. However, we still consider the current approach a preliminary step toward this long-term goal, and recognize significant opportunities for further refinement beyond the present scope.

First, in terms of computational cost, our model is built upon the large variant of DINOv2, which contains approximately 300 million parameters. This results in substantial computational overhead during both training and inference, asking for a certain degree of hardware to fully run the model. To Enhance the accessibility of the model, it requires the adoption of compression techniques such as model pruning or knowledge distillation. However, these approaches come with inherent trade-offs that may affect performance and should be carefully managed.

Second, the current structure of OmniEM processes each downstream task through a separate, independent branch. This modular design ensures flexibility, but limits the potential benefits of multi-task learning. Existing research suggests that learning multiple tasks within a unified architecture can enhance performance and reduce redundancy. However, implementing such a multi-task system introduces challenges: it requires careful architectural design to balance various task demands, and extensive validation to avoid task interference. Furthermore, once a unified structure is established, adding new tasks may become more complex. Given these challenges, we have not yet adopted a multi-task learning approach, but we view it as a promising direction for future work.

References

1. Peddie, C. J. *et al.* Volume electron microscopy. *Nat. Rev. Methods Primers* **2**, 51 (2022).
2. Collinson, L. M. *et al.* Volume em: a quiet revolution takes shape. *Nat. methods* **20**, 777–782 (2023).
3. White, J. G., Southgate, E., Thomson, J. N., Brenner, S. *et al.* The structure of the nervous system of the nematode caenorhabditis elegans. *Philos Trans R Soc Lond B Biol Sci* **314**, 1–340 (1986).
4. Cook, S. J. *et al.* Whole-animal connectomes of both caenorhabditis elegans sexes. *Nature* **571**, 63–71 (2019).
5. Witvliet, D. *et al.* Connectomes across development reveal principles of brain maturation. *Nature* **596**, 257–261 (2021).
6. Zheng, Z. *et al.* A complete electron microscopy volume of the brain of adult drosophila melanogaster. *Cell* **174**, 730–743 (2018).
7. Scheffer, L. K. *et al.* A connectome and analysis of the adult drosophila central brain. *elife* **9**, e57443 (2020).
8. Azevedo, A. *et al.* Connectomic reconstruction of a female drosophila ventral nerve cord. *Nature* **631**, 360–368 (2024).
9. Dorkenwald, S. *et al.* Neuronal wiring diagram of an adult brain. *Nature* **634**, 124–138 (2024).
10. Wanner, A. A., Genoud, C., Masudi, T., Siksou, L. & Friedrich, R. W. Dense em-based reconstruction of the interglomerular projectome in the zebrafish olfactory bulb. *Nat. neuroscience* **19**, 816–825 (2016).
11. Hildebrand, D. G. C. *et al.* Whole-brain serial-section electron microscopy in larval zebrafish. *Nature* **545**, 345–349 (2017).
12. Svara, F. *et al.* Automated synapse-level reconstruction of neural circuits in the larval zebrafish brain. *Nat. Methods* **19**, 1357–1366 (2022).
13. Helmstaedter, M. *et al.* Connectomic reconstruction of the inner plexiform layer in the mouse retina. *Nature* **500**, 168–174 (2013).
14. Kim, J. S. *et al.* Space–time wiring specificity supports direction selectivity in the retina. *Nature* **509**, 331–336 (2014).
15. Ding, H., Smith, R. G., Poleg-Polsky, A., Diamond, J. S. & Briggman, K. L. Species-specific wiring for direction selectivity in the mammalian retina. *Nature* **535**, 105–110 (2016).
16. Morgan, J. L., Berger, D. R., Wetzel, A. W. & Lichtman, J. W. The fuzzy logic of network connectivity in mouse visual thalamus. *Cell* **165**, 192–206 (2016).
17. Turner, N. L. *et al.* Reconstruction of neocortex: Organelles, compartments, cells, circuits, and activity. *Cell* **185**, 1082–1100 (2022).
18. Loomba, S. *et al.* Connectomic comparison of mouse and human cortex. *Science* **377**, eabo0924 (2022).
19. Shapson-Coe, A. *et al.* A petavoxel fragment of human cerebral cortex reconstructed at nanoscale resolution. *Science* **384**, eadk4858 (2024).

20. Xu, C. S. *et al.* An open-access volume electron microscopy atlas of whole cells and tissues. *Nature* **599**, 147–151 (2021).

21. Parlakgül, G. *et al.* Regulation of liver subcellular architecture controls metabolic homeostasis. *Nature* **603**, 736–742 (2022).

22. Miranda, K., Girard-Dias, W., Attias, M., de Souza, W. & Ramos, I. Three dimensional reconstruction by electron microscopy in the life sciences: An introduction for cell and tissue biologists. *Mol. Reproduction Dev.* **82**, 530–547 (2015).

23. Peddie, C. J. & Collinson, L. M. Exploring the third dimension: volume electron microscopy comes of age. *Micron* **61**, 9–19 (2014).

24. Kornfeld, J. & Denk, W. Progress and remaining challenges in high-throughput volume electron microscopy. *Curr. opinion neurobiology* **50**, 261–267 (2018).

25. Poger, D., Yen, L. & Braet, F. Big data in contemporary electron microscopy: challenges and opportunities in data transfer, compute and management. *Histochem. Cell Biol.* **160**, 169–192 (2023).

26. Lee, K., Zung, J., Li, P., Jain, V. & Seung, H. S. Superhuman accuracy on the snemi3d connectomics challenge. *arXiv preprint arXiv:1706.00120* (2017).

27. Buchholz, T.-O. *et al.* Content-aware image restoration for electron microscopy. *Methods cell biology* **152**, 277–289 (2019).

28. Fang, L. *et al.* Deep learning-based point-scanning super-resolution imaging. *Nat. methods* **18**, 406–416 (2021).

29. Lu, C. *et al.* Diffusion-based deep learning method for augmenting ultrastructural imaging and volume electron microscopy. *Nat. Commun.* **15**, 4677 (2024).

30. Yoo, I., Hildebrand, D. G., Tobin, W. F., Lee, W.-C. A. & Jeong, W.-K. ssemnet: Serial-section electron microscopy image registration using a spatial transformer network with learned features. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*, 249–257 (Springer, 2017).

31. Popovych, S. *et al.* Petascale pipeline for precise alignment of images from serial section electron microscopy. *Nat. communications* **15**, 289 (2024).

32. Januszewski, M. *et al.* High-precision automated reconstruction of neurons with flood-filling networks. *Nat. methods* **15**, 605–610 (2018).

33. Sheridan, A. *et al.* Local shape descriptors for neuron segmentation. *Nat. methods* **20**, 295–303 (2023).

34. Heinrich, L. *et al.* Whole-cell organelle segmentation in volume electron microscopy. *Nature* **599**, 141–146 (2021).

35. Conrad, R. & Narayan, K. Instance segmentation of mitochondria in electron microscopy images with a generalist deep learning model trained on a diverse dataset. *Cell Syst.* **14**, 58–71 (2023).

36. Aswath, A., Alsahaf, A., Giepmans, B. N. & Azzopardi, G. Segmentation in large-scale cellular electron microscopy with deep learning: A literature survey. *Med. image analysis* 102920 (2023).

37. Polilov, A. A., Makarova, A. A., Pang, S., Shan Xu, C. & Hess, H. Protocol for preparation of heterogeneous biological samples for 3d electron microscopy: a case study for insects. *Sci. reports* **11**, 4717 (2021).

38. Mulcahy, B. *et al.* A pipeline for volume electron microscopy of the caenorhabditis elegans nervous system. *Front. neural circuits* **12**, 94 (2018).

39. Macrina, T. *et al.* Petascale neural circuit reconstruction: automated methods. *bioRxiv* 2021–08 (2021).

40. Urakubo, H., Bullmann, T., Kubota, Y., Oba, S. & Ishii, S. Uni-em: an environment for deep neural network-based automated segmentation of neuronal electron microscopic images. *Sci. reports* **9**, 19413 (2019).

41. Müller, A. *et al.* Modular segmentation, spatial analysis and visualization of volume electron microscopy datasets. *Nat. Protoc.* **19**, 1436–1466 (2024).

42. Kievits, A. J., Lane, R., Carroll, E. C. & Hoogenboom, J. P. How innovations in methodology offer new prospects for volume electron microscopy. *J. Microsc.* **287**, 114–137 (2022).

43. Belevich, I. & Jokitalo, E. Deepmib: User-friendly and open-source software for training of deep learning network for biological image segmentation. *PLoS computational biology* **17**, e1008374 (2021).

44. Gómez-de Mariscal, E. *et al.* Deepimagej: A user-friendly environment to run deep learning models in imagej. *Nat. methods* **18**, 1192–1195 (2021).

45. Lin, Z., Wei, D., Lichtman, J. & Pfister, H. Pytorch connectomics: a scalable and flexible segmentation framework for em connectomics. *arXiv preprint arXiv:2112.05754* (2021).

46. Patton, W. *et al.* Dacapo: a modular deep learning framework for scalable 3d image segmentation. *arXiv preprint arXiv:2408.02834* (2024).

47. Awais, M. *et al.* Foundation models defining a new era in vision: a survey and outlook. *IEEE Transactions on Pattern Analysis Mach. Intell.* (2025).

48. Tiu, E. *et al.* Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nat. Biomed. Eng.* **6**, 1399–1406 (2022).

49. Zhang, X., Wu, C., Zhang, Y., Xie, W. & Wang, Y. Knowledge-enhanced visual-language pre-training on chest radiology images. *Nat. Commun.* **14**, 4542 (2023).

50. Sun, Y., Wang, L., Li, G., Lin, W. & Wang, L. A foundation model for enhancing magnetic resonance images and downstream segmentation, registration and diagnostic tasks. *Nat. Biomed. Eng.* 1–18 (2024).

51. Zhou, Y. *et al.* A foundation model for generalizable disease detection from retinal images. *Nature* **622**, 156–163 (2023).

52. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T. J. & Zou, J. A visual–language foundation model for pathology image analysis using medical twitter. *Nat. medicine* **29**, 2307–2316 (2023).

53. Chen, R. J. *et al.* Towards a general-purpose foundation model for computational pathology. *Nat. Medicine* **30**, 850–862 (2024).

54. Lu, M. Y. *et al.* A visual-language foundation model for computational pathology. *Nat. Medicine* **30**, 863–874 (2024).

55. Wang, X. *et al.* A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature* **634**, 970–978 (2024).

56. Ma, C., Tan, W., He, R. & Yan, B. Pretraining a foundation model for generalizable fluorescence microscopy-based image restoration. *Nat. Methods* 1–10 (2024).

57. Kirillov, A. *et al.* Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026 (2023).

58. Ma, J. *et al.* Segment anything in medical images. *Nat. Commun.* **15**, 654 (2024).

59. Archit, A. *et al.* Segment anything for microscopy. *Nat. Methods* 1–13 (2025).

60. Shi, R. *et al.* Shapemamba-em: Fine-tuning foundation model with local shape descriptors and mamba blocks for 3d em image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 731–741 (Springer, 2024).

61. Zhuo, Z. *et al.* Segment anything for dendrites from electron microscopy. *arXiv preprint arXiv:2411.02562* (2024).

62. Wan, J. *et al.* Trisam: Tri-plane sam for zero-shot cortical blood vessel segmentation in vem images. *arXiv preprint arXiv:2401.13961* (2024).

63. Mazurowski, M. A. *et al.* Segment anything model for medical image analysis: an experimental study. *Med. Image Analysis* **89**, 102918 (2023).

64. Wang, H. *et al.* Sam-clip: Merging vision foundation models towards semantic and spatial understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3635–3647 (2024).

65. Iudin, A., Korir, P. K., Salavert-Torres, J., Kleywegt, G. J. & Patwardhan, A. Empiar: a public archive for raw electron microscopy image data. *Nat. methods* **13**, 387–388 (2016).

66. Jing, L. & Tian, Y. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis machine intelligence* **43**, 4037–4058 (2020).

67. Perez-Lopez, R., Ghaffari Laleh, N., Mahmood, F. & Kather, J. N. A guide to artificial intelligence for cancer researchers. *Nat. Rev. Cancer* 1–15 (2024).

68. Conrad, R. & Narayan, K. Cem500k, a large-scale heterogeneous unlabeled cellular electron microscopy image dataset for deep learning. *Elife* **10**, e65894 (2021).

69. Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

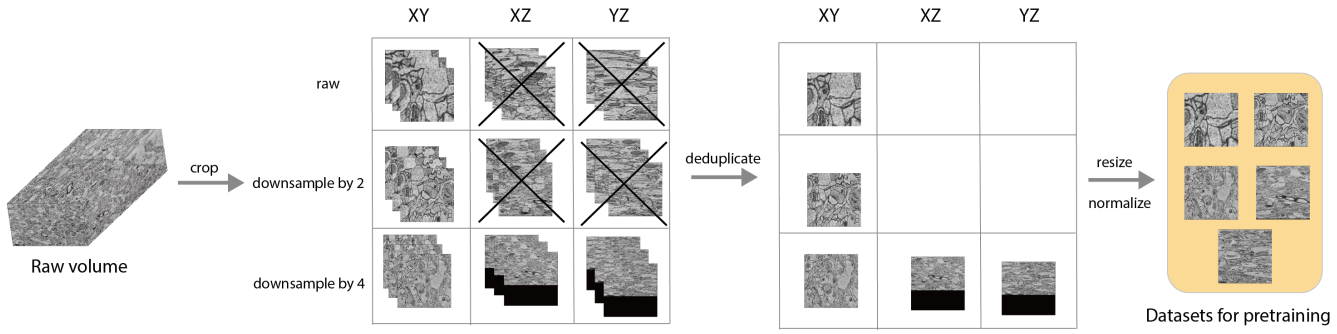
70. Oquab, M. *et al.* Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* (2023).

71. McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

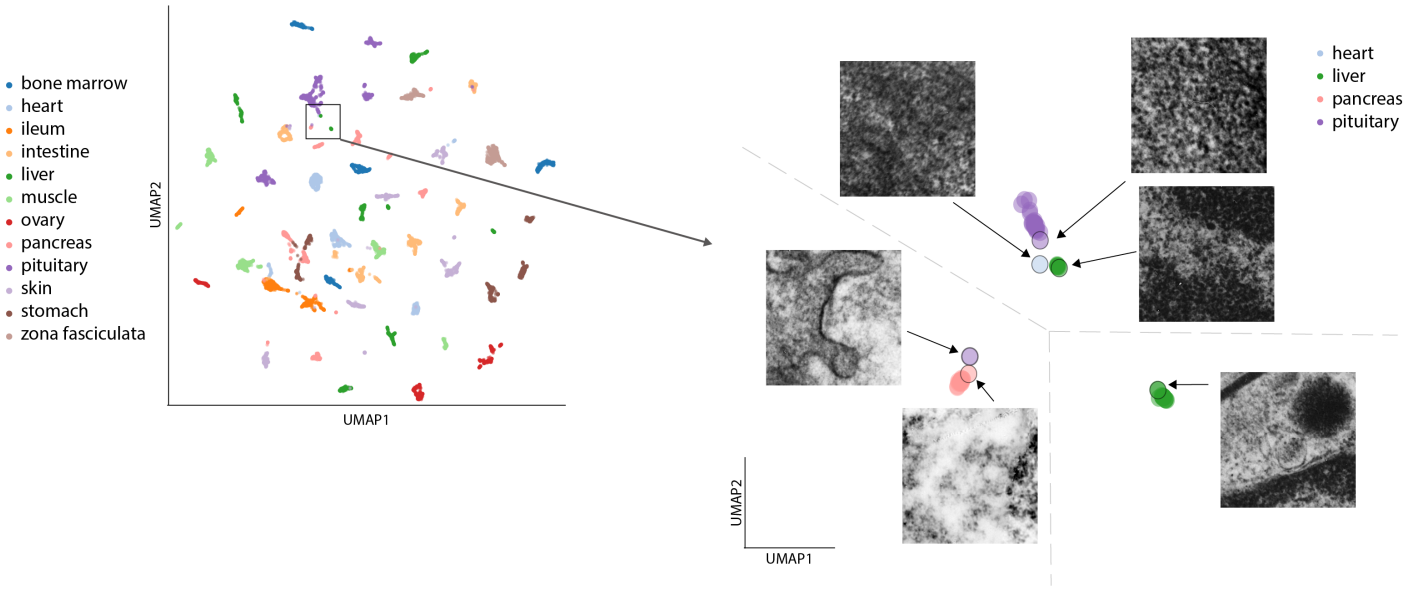
72. Franco-Barranco, D. *et al.* Current progress and challenges in large-scale 3d mitochondria instance segmentation. *IEEE transactions on medical imaging* **42**, 3956–3971 (2023).

73. Müller, A. *et al.* 3d fib-sem reconstruction of microtubule–organelle interaction in whole primary mouse β cells. *J. Cell Biol.* **220**, e202010039 (2020).

Extended data



Extended Data Fig. 1. **EM-5M data curation pipeline**



Extended Data Fig. 2. **EM-DINO distinguishes the organ sources of images based on visual features.**