Unifying the Electron Microscopy Multiverse through a Large-scale Foundation Model

Liuyuan He[1,*], Ruohua Shi[2,*], Wenyao Wang[1,*,+], Yu Cai[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

Accurate analysis of electron microscopy (EM) images is essential for exploring nanoscale biological structures, yet data heterogeneity and fragmented workflows hinder scalable insights. Pretrained on large, diverse datasets, image foundation models provide a robust framework for learning transferable representations across tasks. Here, we introduce EM-DINO, the first image foundational model pretrained on EM-5M, the largest standardized EM corpus (5 million images) encompassing multiple species, tissues, protocols, and resolutions. EM-DINO's multi-scale embeddings capture rich image features that support multiple applications, including organ-specific pattern recognition, image deduplication, and high quality image restoration. Building on these representations, we developed OmniEM, a U-shaped architecture for unified dense prediction that exceeds task-specific models in both image restoration and segmentation. In restoration benchmarks, OmniEM matches the performance of the EM-specific diffusion model while producing fewer hallucinations that could mislead EM interpretation. It also outperforms previous methods in both generalized mitochondrial segmentation and multi-class organelle segmentation. Furthermore, we demonstrate OmniEM's integrated capability to generate high-resolution segmentations from low-resolution inputs, offering the potential to enable fine-scale subcellular analysis in legacy and high-throughput EM datasets. Together, EM-5M, EM-DINO, OmniEM, and an integrated Napari plugin comprise a comprehensive end-to-end toolkit for standardized EM analysis, advance cellular and subcellular understanding and accelerating the discovery of novel organelle morphologies and disease-related alterations.

Main

Electron microscopy (EM) provides an unparalleled window into the nanometer-scale architecture of biological systems, revealing subcellular organelles, synaptic connections, and tissue ultrastructure with extraordinary detail[1–6]. In recent years, high-throughput EM has transformed our ability to visualize cellular and tissue organization at this resolution. While traditional 2D EM relies on expert interpretation of densely packed organelles, volumetric EM (vEM) techniques have emerged over the past two decades to satisfy connectomics' demand for large-scale neuronal reconstructions—generating datasets far too extensive for manual analysis or hypothesis testing[7–9]. Consequently, automated methods are indispensable in EM workflows, spanning low-level tasks (restoration, alignment), segmentation, and higher-order, question-driven analyses such as mapping synaptic connectivity at the circuit level. Unlike fluorescence microscopy, where semantic content derives from labels or contrast agents, EM images are grayscale and rich with nanometer-scale structural information. Therefore, EM research demands bespoke computational models and automated tools to fully harness modern EM's potential and address the pressing biological questions it can illuminate[10].

Modern deep-learning (DL) methods have driven substantial progress in EM image analysis, automating tasks such as segmentation, denoising, and registration[11–16]. Task-specific models now achieve expert-level performance on benchmarks For example, neuronal reconstruction (SNEMI3D, CREMI), organelle identification (MitoNet, BetaSeg), and image restoration (EMDiffuse)[13–15,17]. However, these approaches depend on narrowly defined training sets and often degrade when applied to new imaging conditions or biological contexts[10]. Moreover, EM workflows inherently require distinct models at each processing stage (restoration, alignment, segmentation, and higher-level analysis) imposing significant overhead for tool integration and adaptation[1,18,19], and confining DL solutions to limited task domains (e.g., specialized connectomics pipelines[20]). As a result, existing DL strategies lack the general-purpose flexibility needed for scalable, reproducible EM analysis.

Self-supervised image foundation models (FMs) offer a promising path toward unified, scalable EM workflows. By pretraining on large, diverse unlabeled datasets, FMs learn rich, transferable representations that can be adapted to a wide range of downstream tasks with minimal annotation[21]. In medical imaging, FMs have already shown notable success: pretrained models can be fine-tuned for segmentation, classification, and detection across heterogeneous datasets[22–24]. In EM, however, recent efforts have focused primarily on prompt-based adaptations of the Segment Anything Model (SAM) such as microSAM for general microscopy[25], DendriteSAM for neuronal processes[26], and TriSAM for vascular tracing[27]. These approaches depend on user-provided prompts (e.g., bounding boxes, masks), limiting scalability across the full spectrum of EM tasks from low-level restoration to high-level semantic analysis. Meanwhile, generative models have been applied to restoration[15] but do not learn universally transferable representations. Thus, the EM field still lacks a unified, representation-centric foundation model framework capable of supporting robust, flexible analysis across diverse biological and technical conditions.

To address these gaps, we introduce an end-to-end EM foundation toolkit comprising:

- **EM-5M**: A deduplicated, stratified corpus of 5.5 million 2D EM images drawn from 500 TB of raw data (EMPAIR[28], OpenOrganelle[12], BossDB[29], WebKnossos[30], etc.), spanning seven modalities, 16 species, 30+ organ types, and pixel sizes from 0.5–70 nm.

- **EM-DINO**: A vision transformer pretrained on EM-5M via the DINOv2 self-distillation framework[31], producing robust, multi-scale embeddings that capture shared ultrastructural features.

- **OmniEM**: A U-shaped hybrid network that consumes EM-DINO's final and intermediate patch embeddings and supports universal dense prediction EM tasks. OmniEM delivers state-of-the-art performance in denoising, super-resolution, single-organelle, and multi-organelle segmentation, matching or exceeding diffusion-based models while generating far fewer hallucinations.

- **OmniEM-Napari**: An interactive Napari plugin that packages pretrained EM-DINO and OmniEM checkpoints into a user-friendly GUI, enabling researchers (including those without extensive coding expertise) to deploy OmniEM workflows on their own EM data..

Figure 1 offers a schematic overview of our workflow: large-scale data curation, self-supervised pretraining, and downstream task adaptation. Although foundation-model strategies are well established in natural and medical imaging, EM data pose unique challenges that vEM produces many nearly identical, contiguous slices (< 70 nm apart), so naïvely including adjacent patches would introduce harmful redundancy and imbalance[32]. To mitigate this, we devised a bespoke deduplication pipeline that clusters patches by EM-DINO embeddings or spatial proximity, followed by uninformative filtering (Extended Data Table 1). A clean-train loop, alternating between dataset refinement and EM-DINO checkpoint updates, yields a maximally diverse, nonredundant EM-5M corpus and a robust foundational model(Fig. 1, Left).

Next, we use self-supervised learning to pretrain EM-DINO on EM-5M (Figure 1, Middle Up). In organ classification benchmarks, EM-DINO embeddings formed distinct clusters reflecting organ-specific morphology (Figure 1, Middle Bottom), confirming that large, heterogeneous EM data produce transferable, multi-scale representations. In principle, the EM-DINO embedding space serves as a unified representation supporting low-level processing (alignment, denoising, super-resolution), mid-level understanding (cell and organelle segmentation), and high-level inference (connectomics, spatial organization of subcellular structures).

To operationalize this potential, we developed OmniEM, which translates EM-DINO embeddings into dense predictions across multiple tasks (Figure 1, Right Up). OmniEM's architecture extracts patch-level features from EM-DINO's final and intermediate layers, upsamples them via a multi-scale convolutional decoder, and applies task-specific heads for restoration or segmentation. We demonstrate OmniEM's utility in denoising, super-resolution, mitochondrial segmentation, and multi-class segmentation (Figure 1, Right Middle), where it achieves comparable resolution to diffusion methods while avoiding hallucinations. Finally, we integrated pretrained OmniEM checkpoints into Napari-OmniEM for seamless deployment (Figure 1, Right Bottom).

In the sections that follow, we first evaluate EM-DINO embeddings via organ classification benchmarks (Figure 2). We then assess OmniEM's performance on key downstream tasks—image restoration (Figure 3) and both mitochondrial and multi-organelle segmentation (Figure 4). Lastly, we present practical applications, illustrating how OmniEM's modules can be combined in compound workflows and deployed via the Napari plugin (Figure 5). Collectively, these results highlight the power of EM foundation models to enable scalable, reproducible, end-to-end automation and user-guided refinement in EM analysis. Our full toolkit is publicly available on GitHub.

Results

Evaluating Image Embeddings via Organ Classification

To quantitatively assess the impact of EM-specific pretraining, we designed two organ classification tasks of increasing difficulty to evaluate how well models could distinguish EM images from different organ types and reflect semantic structure in their embeddings. The first, a context-aware task, used images from 11 organ types within *EM-5M* that retained rich morphological cues (Fig.2A). The second, a fine-grained task, employed small field-of-view images from *Histology Guide*[1], which offer limited contextual information and present a more challenging generalization scenario for models (Fig.2D). In the context-aware task, all models exceeded 85% accuracy with just 10% of training data and surpassed 95% with full supervision, with all *EM-DINO* variants outperforming the *DINOv2* baseline (Fig.2B). Zero-shot embedding visualizations further confirmed *EM-DINO*'s advantage, revealing well-separated, semantically coherent clusters compared to the more disorganized *DINOv2* embeddings (Fig.2C). In the more difficult fine-grained task, performance dropped across the board due to reduced context, yet *EM-DINO* maintained over 70% accuracy even with minimal supervision, outperforming all other models including *DINOv2,* which fell below 50%(Fig.2E). With increased training data, *EM-DINO* rapidly surpassed 90% accuracy and converged earlier than its counterparts, while visualizations again highlighted clearer, semantically meaningful clusters (Fig.2F). Additional analysis revealed that *EM-DINO* organized images based on subtle textural and morphological features (Extended Fig.3), indicating its ability to extract fine-grained semantic content. Together, these results demonstrate that self-supervised pretraining on large-scale EM datasets significantly improves the ability to understand and analyze EM images.

Robust Image Restoration with OmniEM

High-quality image restoration is fundamental to EM analysis because high-throughput EM techniques often sacrifice resolution or dwell time (DT) for throughput, precluding reliable detection of fine ultrastructural features. Traditional restoration methods rely on regression, but regression-based models tend to oversmooth images and erase critical nanoscale details. More recently, a generative approach, EMDiffuse[15], has shown promise; however, its diffusion-based process may introduce hallucinated artifacts that mislead downstream analyses. To overcome these challenges, we implemented both denoising and super-resolution within the OmniEM framework. Specifically, we introduced an embedding-based loss derived from EM-DINO's representations, which encourages preservation of fine structural details while mitigating over-smoothing. This enhancement enables OmniEM, a primarily regression-based network, to recover high-resolution features with greater fidelity than existing methods. As a result, OmniEM matches or outperforms state-of-the-art models in both denoising and super-resolution tasks, while offering faster inference and generating significantly fewer hallucinated artifacts (Fig. 3).

To benchmark performance, we compared OmniEM against EMDiffuse and PSSR[33], representing diffusion-based and regression-based approaches, respectively. We adopted the publicly available EMDiffuse dataset, which includes ground truth EM images acquired with long DT (Long-DT) and high resolution (High-R), along with multiple noisy and low-resolution variants captured under short DT (Short-DT) or low resolution (Low-R). Because noise distributions differ significantly across low, medium, and high noise levels, we trained separate regression models (OmniEM and PSSR) for each category, enabling each model to specialize in its specific noise characteristics.

Notably, EM-DINO's embedding-based loss contributes substantially to these gains (Extended Data Fig. 4; see Methods). We introduce two hyperparameters, α for image-level embedding loss and β for patch-level embedding loss, to control its influence during training. Incorporating the embedding loss markedly reduces unwanted smoothness in the denoised output, especially at high noise levels (Extended Data Fig. 4B). For example, under high-noise conditions, LPIPS on denoised images drops from 0.52 (without embedding loss) to 0.18 (with $\alpha = 0.25$, $\beta = 0$), indicating significantly improved perceptual fidelity. Varying α and β across a reasonable range does not significantly change LPIPS; however, larger embedding-loss weights can introduce spurious patterns that LPIPS fails to penalize. Because current metrics cannot distinguish these artifacts, we chose $\alpha = 0.25$ and $\beta = 0$ to balance reduced smoothness against minimal pattern hallucination in all subsequent experiments.

In the denoising task, OmniEM balances detail preservation and artifact suppression across all noise levels (Fig. 3A–D). Under low-noise conditions, all methods produce high-quality outputs: PSSR maintains membrane boundaries but can appear over-smoothed, while EMDiffuse and OmniEM both sharpen ultrastructural features (Fig. 3A). Quantitatively (Fig. 3B), OmniEM matches EMDiffuse in PSNR, FSIM, and FRC resolution (pairwise ratios cluster around 1) but surpasses both EMDiffuse and PSSR in LPIPS, indicating superior perceptual and structural fidelity to long-DT images. Under high-noise conditions (Fig. 3C), differences become more pronounced: PSSR's stronger smoothing preserves general morphology but misses finer details; EMDiffuse recovers textures that are visually plausible but include spurious artifacts not found in the long-DT ground truth; in contrast, OmniEM conservatively restores key cellular topologies without hallucination, preserving structural consistency essential for downstream tasks such as neuron segmentation and vesicle counting (Extended Data Fig. 5A,B). Quantitatively, OmniEM's PSNR falls between EMDiffuse and PSSR, while its FSIM exceeds both; in LPIPS

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

and FRC resolution, OmniEM slightly outperforms EMDiffuse and markedly outperforms PSSR (Fig. 3D). Results under medium-noise conditions follow the same pattern (Extended Data Fig. 6A). Overall, these findings demonstrate that OmniEM's embedding-based loss underpins its robust performance across noise regimes.

Beyond denoising, OmniEM also excels in super-resolution (Fig. 3E–H), effectively narrowing the gap between low- and high-resolution EM imaging. Its super-resolution module reconstructs critical subcellular details from low-resolution inputs under high noise (Fig. 3E), low noise (Fig. 3G), and medium noise (Extended Data Fig. 6B). In contrast, PSSR yields overly smoothed results while EMDiffuse recovers fine details but often introduces artificial textures absent from the high-resolution reference (Fig. 3G). OmniEM strikes a balanced compromise: it sharpens ultrastructure without hallucinating spurious features, a quality essential for downstream analyses such as vesicle counting (Extended Data Fig. 5C). Quantitatively (Fig. 3F,H), OmniEM matches EMDiffuse in LPIPS and FRC resolution, both of which emphasize perceptual and spatial accuracy, and significantly outperforms PSSR. Moreover, OmniEM slightly surpasses EMDiffuse in PSNR and FSIM, indicating superior pixel-wise fidelity and feature-based similarity.

In conclusion, OmniEM delivers image restoration performance on par with state-of-the-art methods while substantially reducing hallucinated structures, especially under high noise, by producing more conservative and biologically plausible reconstructions. This reliability is critical for downstream tasks that depend on accurate ultrastructural representation. Further improvements may be achieved by exploring enhanced architectural designs or loss functions.

Precious image segmentation

The impressive performance of OmniEM in image restoration reflects its robust modeling of the statistical properties of EM data, we then ask whether OmniEM's embeddings, learned through EM-DINO, support higher level understanding of cellular ultrastructure. To this end, we evaluated the model on organelle segmentation tasks, a longstanding critical step in EM analysis. Accurate organelle segmentation forms the basis for quantitative ultrastructural analysis and biological discovery, while existing methods struggle to generalize across diverse EM datasets and semantic categories. Through a zero-shot mitochondrial segmentation task and a multi-class segmentation task encompassing mitochondria, Golgi apparatus, secretory granules, and nuclei, we demonstrated OmniEM's strong generalization capability and its efficient adaptability. Again, these findings highlight the model's potential as a general-purpose tool for EM image interpretation.

Mitochondrial morphology and distribution serve as critical indicators of cellular metabolic state and disease progression. However, the substantial morphological heterogeneity of mitochondria across tissues poses a major challenge for cross-dataset generalization in automated segmentation models[34]. To evaluate OmniEM's ability to overcome this challenge, we fine-tuned it on the CEM-MitoLab dataset and compared its performance to MitoNet on seven external benchmark datasets (all withheld from training) following the protocol of Conrad et al.[13]. As expected, the publicly released MitoNet checkpoint maintains strong segmentation accuracy across all seven benchmarks (Figure 4A, third row; light-apricot histogram bars in Figure 4B). OmniEM, however, matches or exceeds MitoNet's performance on five of seven datasets (Figure 4A, below row; light-blue histogram bars in Figure 4B)), highlighting the value of EM-DINO pretraining in providing robust, transferable embeddings for mitochondria segmentation. Overall, OmniEM raises the mean Dice score from 61.8 % (MitoNet) to 66.0 % (Figure 4B). The largest improvement occurs on the salivary gland dataset, where mitochondria exhibit flat, bowl-shaped geometries, high organelle density, and weak contrast against salivary granules (Figure 4A, sixth column). In this particularly challenging case, OmniEM's Dice score leaps from 11.3 % (MitoNet) to 30.7 % (Figure 4B, sixth inset), demonstrating that EM-DINO pretraining confers enhanced representational capacity for segmenting rare or low-contrast mitochondrial morphologies. These results underscore OmniEM's superior generalization and zero-shot performance, even on unseen tissue types and diverse imaging conditions.

Compared to single-organelle segmentation, multi-organelle segmentation is considerably more challenging due to the diverse morphologies, sizes, and close spatial relationships among different organelles. Variability in imaging conditions, staining quality, and the inherent ultrastructural complexity further complicate accurate delineation, often resulting in overlapping boundaries and ambiguous separations between adjacent structures. To address these challenges, we implemented a multi-class segmentation head within OmniEM and evaluated its performance on the BetaSeg dataset[17], which includes annotations for Golgi apparatus, mitochondria, insulin secretory granules, and nuclei (all other pixels are treated as background). For baseline comparison, we selected GobletNet[35], a recently published model that uses wavelet transforms to extract high-frequency features and a fusion-attention module to improve segmentation accuracy across EM datasets. We followed its experimental protocol by subsampling the full BetaSeg training set (30 % of the total) into four fractions, 1 %, 30 %, 60 %, and 100 %, to test few-shot learning capability.

As shown in Figure 4C and Extended Data Figure 7, OmniEM consistently outperformed GobletNet across most training fractions. Remarkably, with only 1 % of the training data, OmniEM achieves a higher overall mean Dice score without relying on explicit high-frequency priors. In particular, OmniEM's Golgi segmentation improves from 28.0 % (GobletNet) to 63.8 % Dice (Figure 4C, second column, blue masks; Figure 4D, top right inset). GobletNet outperforms OmniEM on mitochondrial

and nuclear segmentation at the 1 % training fraction, likely because its wavelet-driven features are sensitive to sharp nuclear boundaries. However, GobletNet suffers from high false positives, as it cannot reliably distinguish nuclear from non-nuclear pixels (Extended Data Figure 8). Beyond these exceptions, OmniEM exceeds GobletNet's performance for all organelle classes under every training regime. Overall, OmniEM raises the mean Dice from 88.7 % to 94.1 %, with class-wise gains from 72.1 % to 83.0 % for Golgi, 87.1 % to 94.0 % for insulin secretory granules, 92.3 % to 97.1 % for mitochondria, and 99.2 % to 99.7 % for nuclei.

These results highlight OmniEM's strong generalization and data efficiency, enabled by the rich representational capacity inherited from EM-DINO pretraining. Without relying on handcrafted features or task-specific priors, OmniEM effectively captures diverse organelle morphologies and subtle ultrastructural cues. This underscores the potential of EM foundation models to deliver robust, versatile segmentation across complex electron microscopy datasets.

Integrating OmniEM into Scalable EM Workflows

OmniEM's multi-task design enables streamlined, end-to-end EM workflows by combining advanced restoration, super-resolution, and segmentation within a single, user-friendly framework (Fig. 5).

By chaining OmniEM's super-resolution and segmentation branches, researchers can execute complex pipelines without switching tools. In the glycolytic-muscle benchmark, we downsampled images to simulate low-resolution inputs and compared three scenarios: (1) segmentation on original-resolution images (18 nm pixel size), (2) segmentation after bilinear upsampling of low-resolution inputs (36 nm pixel size), and (3) segmentation after applying OmniEM's super-resolution to the same low-resolution images (36 nm pixel size). OmniEM's super-resolution output (LR-OmniEM) visibly restores fine mitochondrial morphology lost in bilinear interpolation, recovering cristae and membrane contours that are otherwise indistinct (Fig. 5A). When these super-resolved images are fed into OmniEM's segmentation branch, previously missed mitochondria are successfully identified (Fig. 5B, right), whereas bilinear upsampling alone fails to reveal many of these structures (Fig. 5B, center). Three-dimensional renderings of the segmentation confirm that volumes derived from super-resolved data closely resemble the ground-truth ultrastructure, preserving tubular shapes and interconnections (Fig. 5C, right). Quantitatively, segmentation on super-resolved images achieves a Dice coefficient of 61.69%, indicating substantial overlap with the ground truth, while segmentation on bilinearly upsampled, low-resolution inputs yields only 16.36% Dice (Fig. 5C, center). By comparison, segmentation on the original high-resolution images delivers a Dice score of 79.96%. Taken together, these results suggest that OmniEM's super-resolution can substantially bridge the gap between low-resolution acquisition and high-accuracy segmentation, although there remains room for improvement to approach the performance achieved on native high-resolution data.

We have packaged OmniEM's restoration, super-resolution, and segmentation modules into an interactive Napari plugin (Fig. 5D), allowing users to apply OmniEM via a graphical interface. Users can load images, adjust parameters, and visualize results in real time—no coding required. By bundling pretrained EM-DINO and OmniEM checkpoints within this plugin, we provide a comprehensive, platform-based solution for scalable, high-quality EM analysis with minimal technical overhead.

Discussion

In this work, we demonstrate that large-scale, EM-specific self-supervised learning enables the construction of a foundational vision model capable of unifying previously fragmented EM workflows. At the core of our approach is a novel data-curation pipeline that systematically deduplicates and filters vEM patches, reducing redundancy while keeping biological and technical diversity. By pairing EM-DINO for robust feature extraction with OmniEM for flexible downstream prediction tasks, we deliver a suite of tools tailored to the EM community's demand for efficient, scalable, and reproducible analysis. Although achieving state-of-the-art benchmarks was not our primary objective, OmniEM nonetheless attains strong performance in denoising, super-resolution, and segmentation without reliance on handcrafted heuristics or specialized architectures, highlighting the generalization power of EM-DINO's learned representations.

At its foundation, representation learning is the process of extracting meaningful, structured information from raw data to support further understanding and analysis[36]. Through multiple use cases, including large-scale data curation, embedding-based loss design, and diverse dense prediction tasks, we show that EM-DINO produces embeddings that are both transferable and reliable across applications. Because most modern image processing and interpretation pipelines hinge on the availability of suitable features, EM-DINO can serve as a universal feature extractor, readily integrated into existing tools to enhance performance. For example, in image stitching and alignment, point correspondence detection often depends on classical descriptors such as SIFT[37], or more recently, deep encoders trained to suppress artifacts in EM images[16,18]. EM-DINO's semantically enriched features could significantly improve matching accuracy across sections, particularly under varying imaging conditions. Similarly, in connectomics pipelines that require the linking of neurite fragments, cellular identity assignment, and synapse detection, practitioners typically rely on separate classifiers optimized for each task[38]. We propose

that EM-DINO can serve as a unified representation backbone across these components, reducing task fragmentation and simplifying pipeline integration.

Image enhancement and segmentation are essential tasks in EM analysis, both requiring pixel-level understanding of complex biological structures. OmniEM is, to our knowledge, the first unified framework to address both tasks within a single architecture by sharing a common backbone and applying task-specific output heads. This design enables integrated workflows where restoration directly improves segmentation, especially by enhancing organelle boundaries and reducing image noise. The utility of this joint approach has been demonstrated in Cellpose 3 for light microscopy[39], but it holds even greater promise in EM. By enabling accurate segmentation from low-resolution or undersampled images, OmniEM offers a practical path to reduce imaging demands without sacrificing output quality. This capability is particularly valuable in large-scale EM studies, where throughput and cost are limiting factors. Previous models have shown that restoration enhances segmentation, but often rely on manually annotated data or isolated pipelines[15,33]. In contrast, OmniEM provides an end-to-end system deployable via a Napari plugin, allowing researchers to test real-world low-quality inputs directly. This integration of restoration and segmentation into a unified, accessible tool marks a significant advance toward scalable, generalizable EM analysis.

While our work establishes a robust EM-specific foundation, several avenues can further improve generalization and efficiency. First, expanding EM-5M beyond its current emphasis on brain and mouse vEM to include underrepresented organs (e.g., kidney, heart, bone) and non-mammalian species will ensure downstream models learn truly transferable representations across diverse EM applications. Second, although EM-DINO uses a Vision Transformer backbone, alternative architectures, such as hierarchical transformers like Swin Transformer[40] or Mamba[41], may better capture EM's anisotropic resolution and high noise levels. Systematic benchmarking of these backbones on EM-specific tasks will clarify which design choices maximize ultrastructural detail. Finally, EM-DINO's large-scale pretraining yields robust embeddings, but fine-tuning remains computationally intense; parameter-efficient techniques like Low-Rank Adaptation (LoRA)[42] offer a promising route to rapid adaptation on new datasets with minimal overhead. Moreover, modern EM workflows increasingly rely on 3D volumes rather than isolated 2D slices. While EM-DINO's 2D embeddings serve many tasks, they cannot fully encode volumetric context. Future work should develop models that natively process 3D volumes by extracting multi-scale representations from isotropic or anisotropic voxel grids via spatial-temporal adapters[43] or related mechanisms, so that OmniEM can handle real-world 3D datasets (e.g., the CellMap whole-cell segmentation challenge[2]) more effectively. Additionally, OmniEM currently employs separate branches for restoration and segmentation, missing potential synergies from shared features. A multi-task or auxiliary-learning framework could unify these tasks, improving both efficiency and accuracy[44], and moving closer to a truly general-purpose, dense-prediction model for EM data. By addressing these challenges, future work can build on our EM-specific foundation to create ever more versatile, accurate, and efficient tools for nanoscale biological discovery.

References

1. Loomba, S. *et al.* Connectomic comparison of mouse and human cortex. *Science* **377**, eabo0924 (2022).
2. Xu, C. S. *et al.* An open-access volume electron microscopy atlas of whole cells and tissues. *Nature* **599**, 147–151 (2021).
3. Witvliet, D. *et al.* Connectomes across development reveal principles of brain maturation. *Nature* **596**, 257–261 (2021).
4. Svara, F. *et al.* Automated synapse-level reconstruction of neural circuits in the larval zebrafish brain. *Nat. Methods* **19**, 1357–1366 (2022).
5. Parlakgül, G. *et al.* Regulation of liver subcellular architecture controls metabolic homeostasis. *Nature* **603**, 736–742 (2022).
6. Shapson-Coe, A. *et al.* A petavoxel fragment of human cerebral cortex reconstructed at nanoscale resolution. *Science* **384**, eadk4858 (2024).
7. Peddie, C. J. & Collinson, L. M. Exploring the third dimension: volume electron microscopy comes of age. *Micron* **61**, 9–19 (2014).
8. Peddie, C. J. *et al.* Volume electron microscopy. *Nat. Rev. Methods Primers* **2**, 51 (2022).
9. Collinson, L. M. *et al.* Volume em: a quiet revolution takes shape. *Nat. methods* **20**, 777–782 (2023).
10. Kornfeld, J. & Denk, W. Progress and remaining challenges in high-throughput volume electron microscopy. *Curr. opinion neurobiology* **50**, 261–267 (2018).
11. Buchholz, T.-O. *et al.* Content-aware image restoration for electron microscopy. *Methods cell biology* **152**, 277–289 (2019).
12. Heinrich, L. *et al.* Whole-cell organelle segmentation in volume electron microscopy. *Nature* **599**, 141–146 (2021).

[2] https://cellmapchallenge.janelia.org/

13. 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).

14. 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).

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

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

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

18. Bae, J. A. *et al.* Functional connectomics spanning multiple areas of mouse visual cortex. *Nature* **640**, 435–447, DOI: 10.1038/s41586-025-08790-w (2025).

19. 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).

20. Dorkenwald, S. *et al.* Cave: Connectome annotation versioning engine. *Nat. Methods* DOI: 10.1038/s41592-024-02426-z (2025).

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

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

23. 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).

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

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

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

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

28. Iudin, A. *et al.* Empiar: the electron microscopy public image archive. *Nucleic Acids Res.* **51**, D1503–D1511, DOI: 10.1093/nar/gkac1062 (2022).

29. Hider, R. *et al.* The brain observatory storage service and database (bossdb): A cloud-native approach for petascale neuroscience discovery. *Front. Neuroinformatics* **Volume 16 - 2022**, DOI: 10.3389/fninf.2022.828787 (2022).

30. Boergens, K. M. *et al.* webknossos: efficient online 3d data annotation for connectomics. *Nat. methods* **14**, 691–694 (2017).

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

32. Vo, H. V. *et al.* Automatic data curation for self-supervised learning: A clustering-based approach. *arXiv preprint arXiv:2405.15613* (2024).

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

34. 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).

35. Zhou, Y., Li, L., Wang, C., Song, L. & Yang, G. Gobletnet: Wavelet-based high-frequency fusion network for semantic segmentation of electron microscopy images. *IEEE Transactions on Med. Imaging* (2024).

36. Le-Khac, P. H., Healy, G. & Smeaton, A. F. Contrastive representation learning: A framework and review. *Ieee Access* **8**, 193907–193934 (2020).

37. Lowe, D. G. Distinctive image features from scale-invariant keypoints. *Int. journal computer vision* **60**, 91–110 (2004).

38. Motta, A. *et al.* Dense connectomic reconstruction in layer 4 of the somatosensory cortex. *Science* **366**, eaay3134 (2019).

39. Stringer, C. & Pachitariu, M. Cellpose3: one-click image restoration for improved cellular segmentation. *Nat. Methods* 1–8 (2025).

40. Liu, Z. *et al.* Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022 (2021).

41. Liu, X., Zhang, C. & Zhang, L. Vision mamba: A comprehensive survey and taxonomy. *arXiv preprint arXiv:2405.04404* (2024).

42. Hu, E. J. *et al.* Lora: Low-rank adaptation of large language models. *ICLR* **1**, 3 (2022).

43. Pan, J., Lin, Z., Zhu, X., Shao, J. & Li, H. St-adapter: Parameter-efficient image-to-video transfer learning. *Adv. Neural Inf. Process. Syst.* **35**, 26462–26477 (2022).

44. Yu, J. *et al.* Unleashing the power of multi-task learning: A comprehensive survey spanning traditional, deep, and pretrained foundation model eras. *arXiv preprint arXiv:2404.18961* (2024).

Acknowledgements

We thank Quanyue Ma for assistance in downloading EM images from HistologyGuide.

Author contributions statement

These authors contributed equally: Liuyuan He, Ruohua Shi, and Wenyao Wang.

Contributions

L.M. and W.W. supervised the research. L.M., W.W. and L.H. conceived of the technique. W.W., L.H., R.S., and Y.C. implemented the algorithm and conducted the experiments. L.H. curated the pretraining dataset, pretrained backbone models, designed the framework of OmniEM, and developed OmniEM-Napari. R.S. implemented the downstream tasks and organized the codes and models. All authors had access to the study and wrote the paper.

Corresponding author

Correspondence to Lei Ma.

Figures

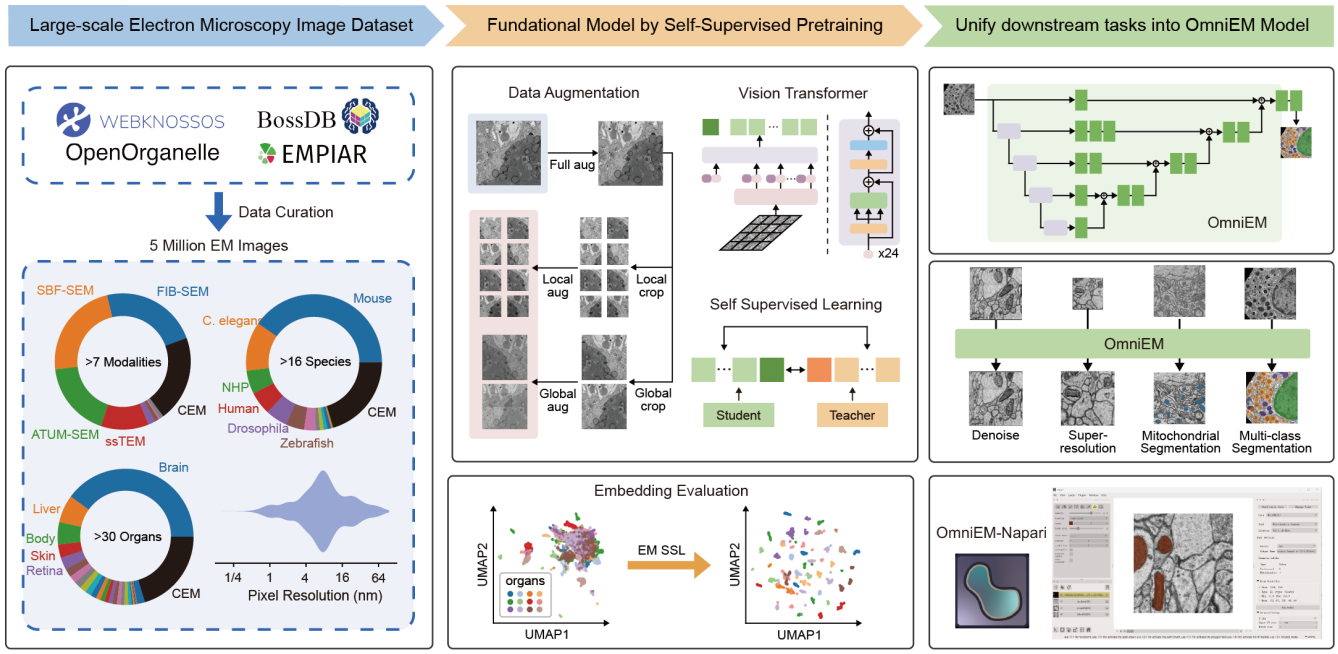


Fig. 1. Overview of *EM-5M*, *EM-DINO*, and *OmniEM*. Left. Data from Webknossos, BossDB, OpenOrganelle, EMPIAR and other sources are collected and curated to *EM-5M*, which contains over 5 million standardized images. The EM-5M spectrum spans over seven imaging devices, more than 16 animal species, over 30 organ types, and a broad range of pixel sizes from 0.5 nm to 70 nm. **Middle.** Illustration of the SSL process of *EM-DINO*. EM-specific data augmentations were introduced into the SSL process. After SSL with the large-scale *EM-5M* dataset, *EM-DINO* can form semantically meaningful embeddings from EM images. **Right.** From top to bottom: the framework of OmniEM, supported downstream tasks, and the interface of the Napari plugin.

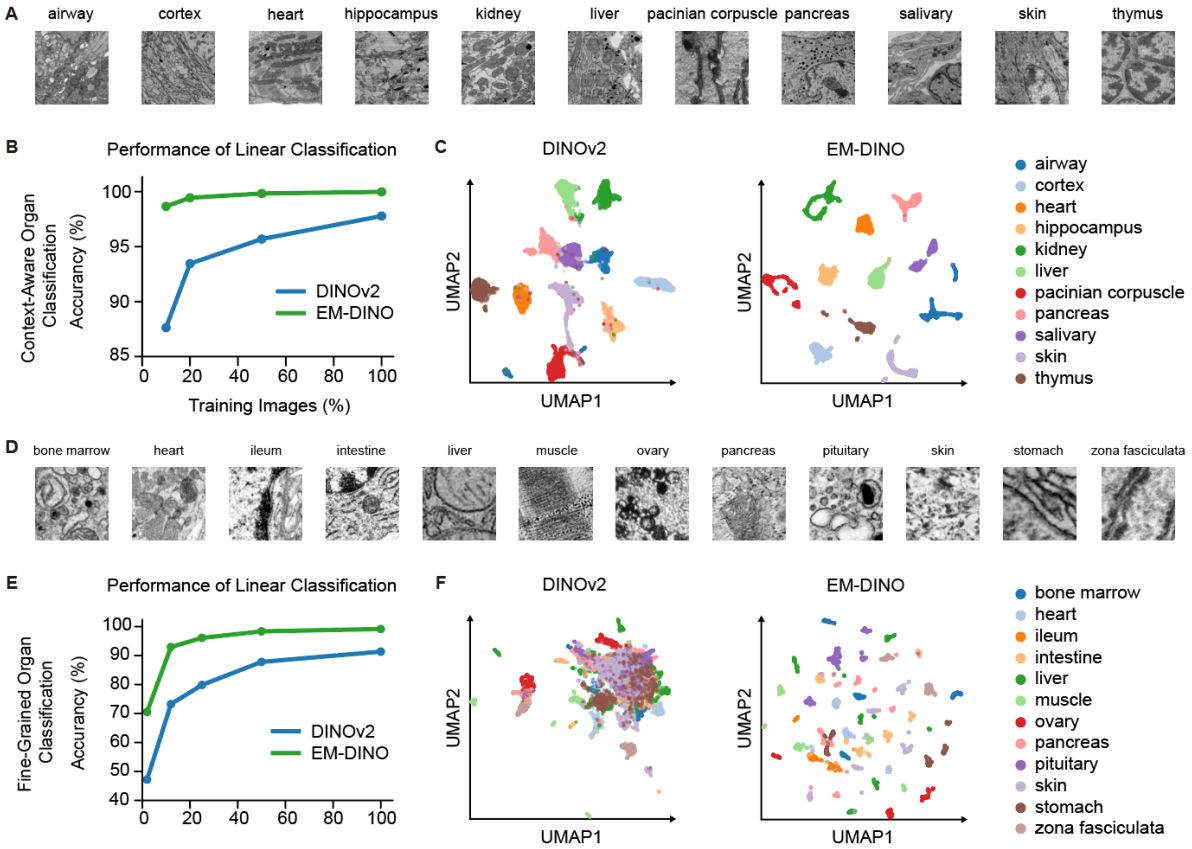


Fig. 2. Evaluation of model performance on organ classification tasks. The *EM-DINO* model, pretrained with diverse EM-specific data, demonstrates superior representational capability compared to the natural image–pretrained *DINOv2* through self-supervised learning. (**A**) Schematic of the context-aware classification task using morphologically rich EM image patches from 11 organ types. (**B**) Classification accuracy of *DINOv2* and *EM-DINO* across varying amounts of training data. (**C**) Zero-shot UMAP visualization of image embeddings from the context-aware task, showing clearer organ-level separation with *EM-DINO*. (**D**) Schematic of the fine-grained classification task using small field-of-view, low-context EM images from 12 organs in an out-of-distribution dataset. (**E**) Classification accuracy of both models on the fine-grained task across different training data sizes. (**F**) Zero-shot UMAP visualization of embedding spaces in the fine-grained setting, where *EM-DINO* forms more distinct and semantically meaningful clusters.

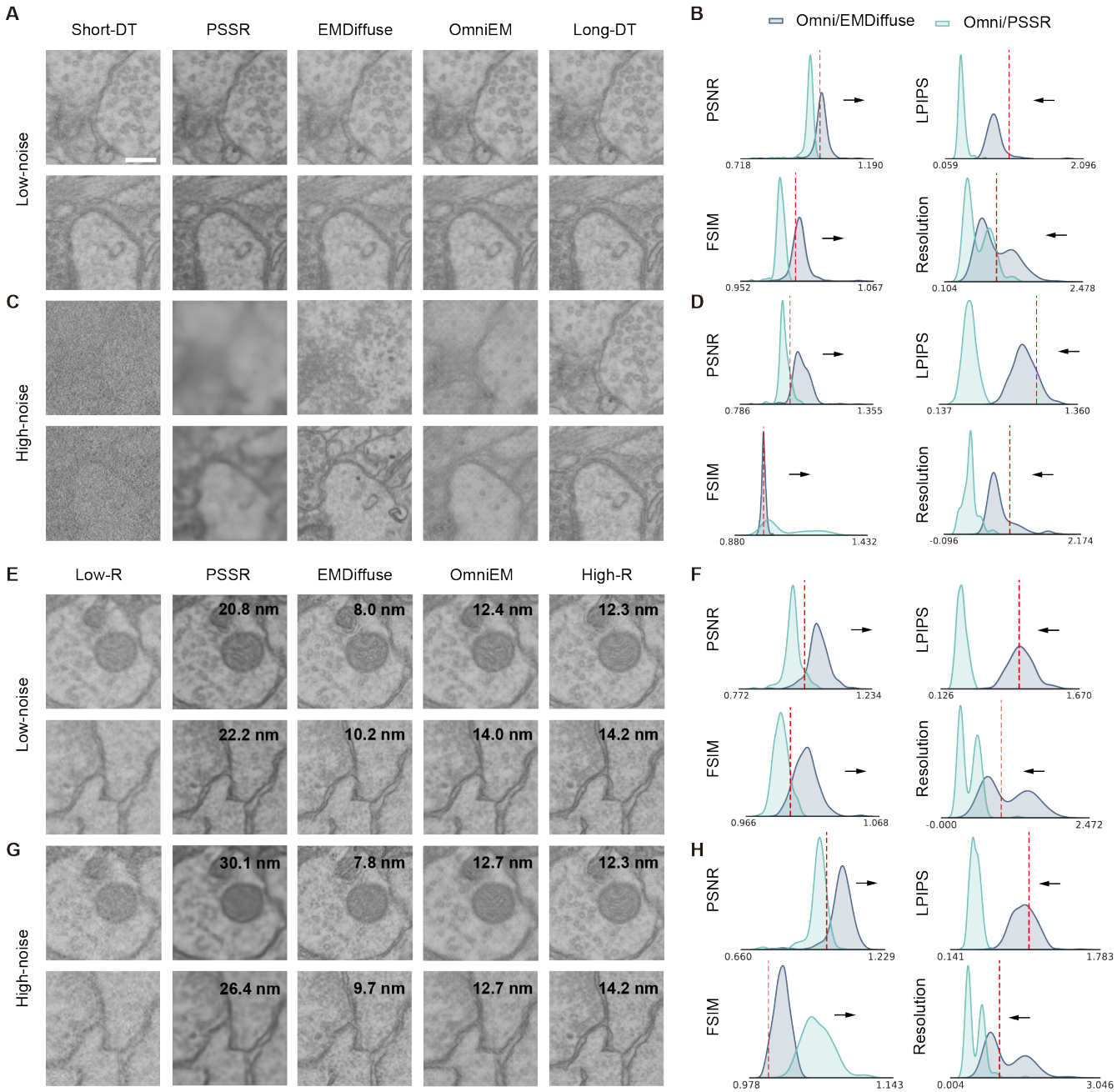


Fig. 3. Evaluation of OmniEM Image Denoising and Super-Resolution. (A) Comparison of the denoising results from different models under low noise conditions. From left to right: the noisy image acquired with short dwell time (used as model input), the restored results from PSSR, EMDiffuse, and OmniEM, and the high-quality image acquired with long dwell time (used as ground truth reference). Two image regions are shown as representative cases. **(B)** Shown are the relative performance of different models on all pair-wise test data under low noise levels, using four metrics: PSNR, LPIPS, FSIM, and resolution. Light blue line represents the ratio distribution of OmniEM to EMDiffuse, and light green line represents the ratio distribution of OmniEM to PSSR. The vertical red dashed line indicates the ratio value of 1, and arrows denote the direction of better performance for each metric. **(C and D)** show the corresponding qualitative and quantitative denoising results under high noise conditions, with the same layout and legend conventions as in A and B, respectively. **(E)** Comparison of the super-resolution results from different models under low noise conditions. From left to right: the low-resolution input image, super-resolved outputs from PSSR, EMDiffuse, and OmniEM, and the high-resolution ground truth image. Two representative regions are shown. The resolution of each restored image is displayed in the upper right corner. **(F)** Shown are the relative performance of different models on all pairwise test data under low noise conditions for super-resolution, using the same four metrics. **(G and H)** show the corresponding qualitative and quantitative super-resolution results under high noise conditions, following the same layout and conventions as in E and F. Scale bar: 200 nm.

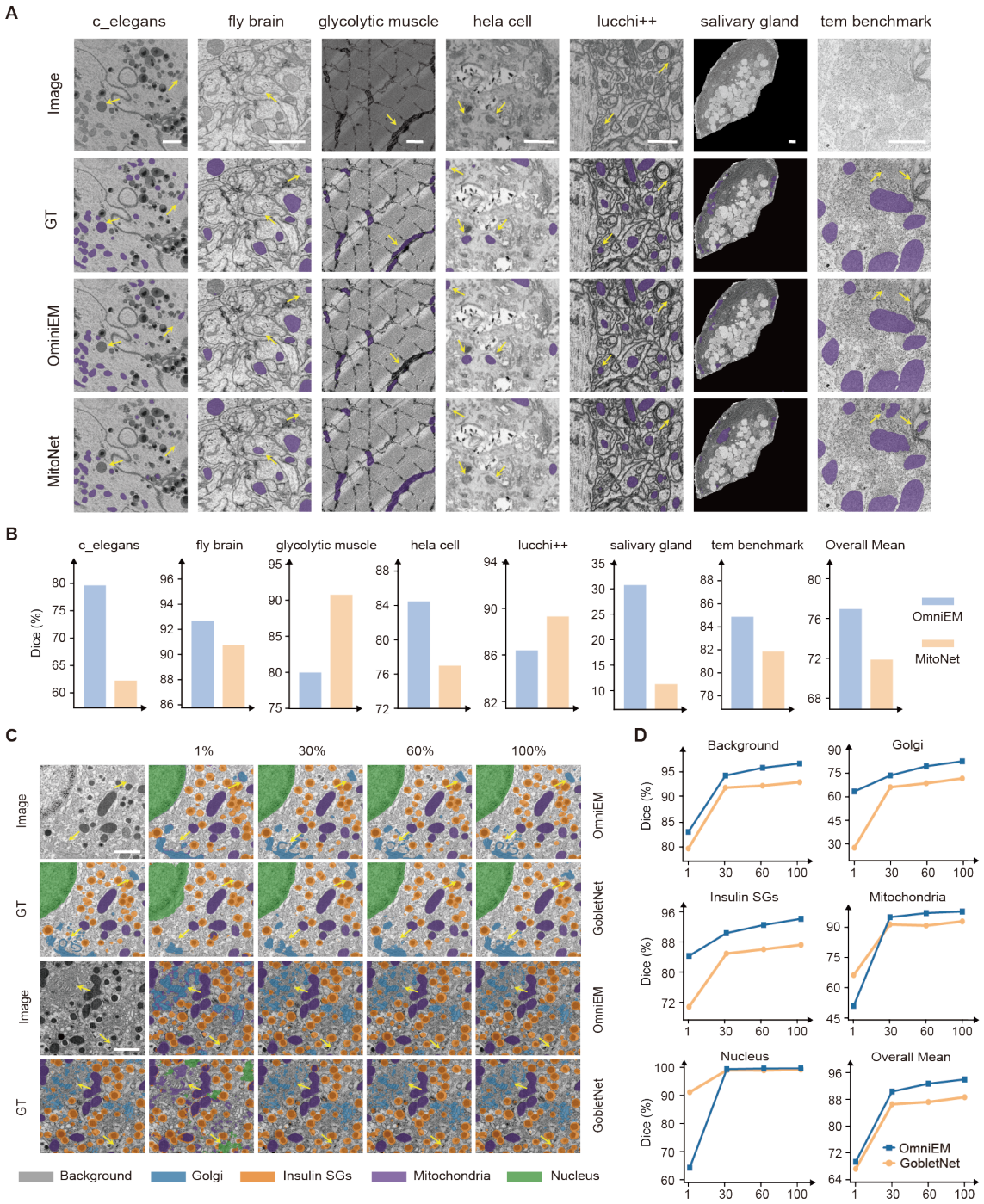


Fig. 4. Segmentation for Mitochondrail and Multi-organelle. (**A**) Mitochondrial segmentation outputs on seven datasets, from left to right: C. elegans, fly brain, glycolytic muscle, HeLa cell, Lucchi++, salivary gland, and TEM images benchmark. From top to bottom: raw input, ground truth (GT), OmniEM predictions and MitoNet predictions. Purple masks represent GT labels or predicted segmentations overlaid on the raw input images. Yellow arrows highlight regions where OmniEM precisely traces mitochondrial boundaries, whereas MitoNet tends to over-segment or miss these structures. **(B)** Dice coefficient comparisons between OmniEM (light blue) and MitoNet (light apricot) across the corresponding seven benchmark datasets in A; the final inset summarizes the mean Dice scores across all datasets. (**C**) Representative EM images and segmentation results on Betaseg dataset. Column one, raw input images and GT labeling. Column two to five, segmentation results at the training ratio of 1%, 30%, 60%, and 100%. Odd row, segmentation results of OmniEM. Even row, segmentation results of GobletNet. Yellow arrows indicate regions where OmniEM precisely segments while GobletNet over-segments or misses. Colors: background, grey; Golgi apparatus, blue; insulin SGs, orange; mitochondria, purple; nucleus, green. (**D**) Class-specific Dice coefficient curves for OmniEM (blue) and GobletNet (light apricot) across training ratios. Scale bar, 1 μm, applies to all insets.

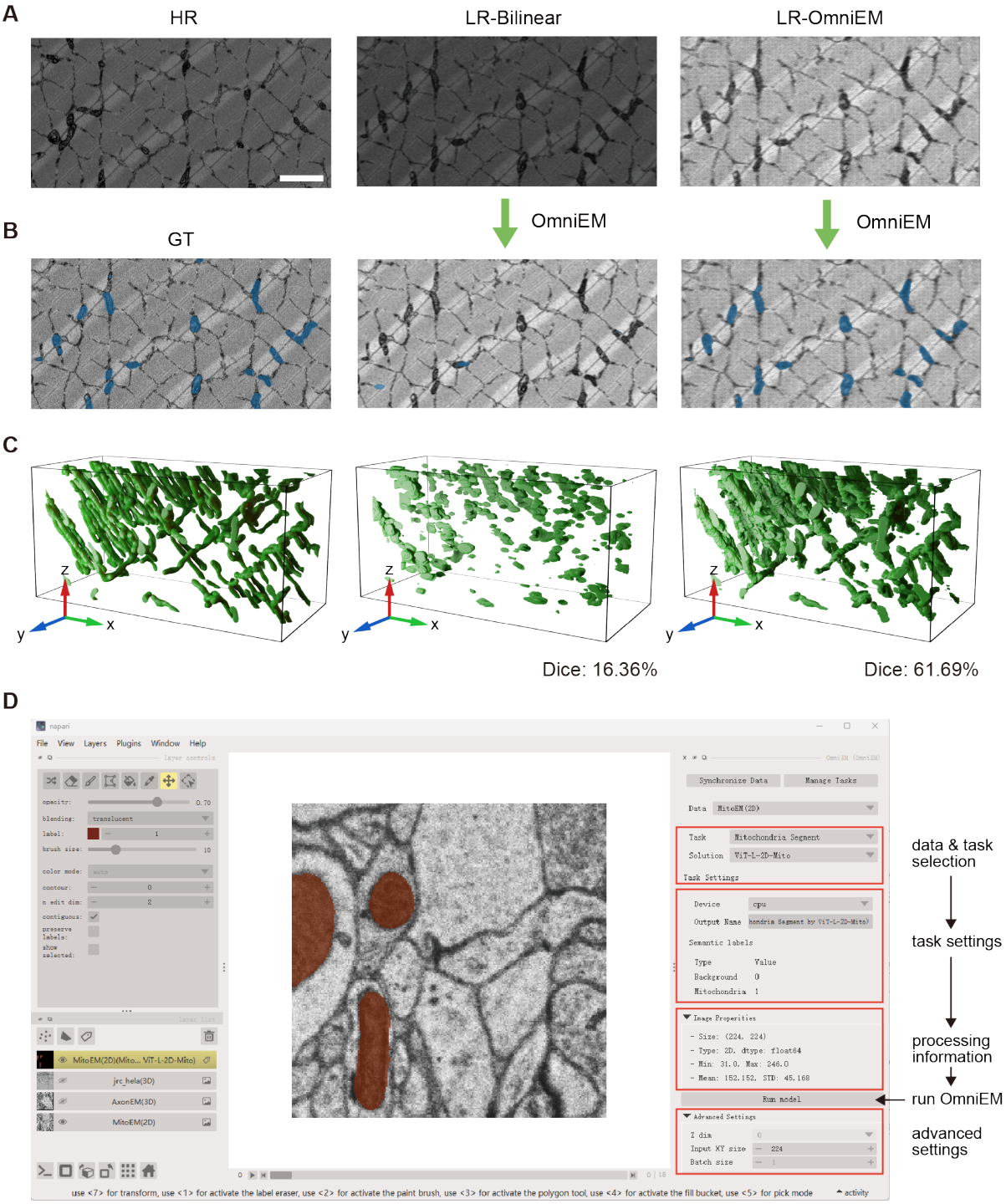


Fig. 5. Empowering EM Image Analysis with OmniEM and Integrated Napari Plugin (**A**) From left to right: original high-resolution image from the glycolytic-muscle mitochondrial segmentation benchmark; low-resolution image upsampled by bilinear interpolation (LR-Bilinear); low-resolution image enhanced by OmniEM super-resolution (LR-OmniEM). (**B**) From left to right: ground-truth mitochondrial mask; segmentation on LR-Bilinear input; segmentation on LR-OmniEM input. (**C**) 3D reconstructions of mitochondria from: original high-resolution segmentation (left), LR-Bilinear segmentation (center), and LR-OmniEM segmentation (right). (D) Screenshot of the OmniEM-Napari plugin interface. Scale bar: 2 μm.

Extended data

Source Data 1: Performance comparison of Organ Classification Tasks (Accuracy in %)

Source Data 2: values of metrics in figure 3

Source Data 3: values of metrics in figure 4

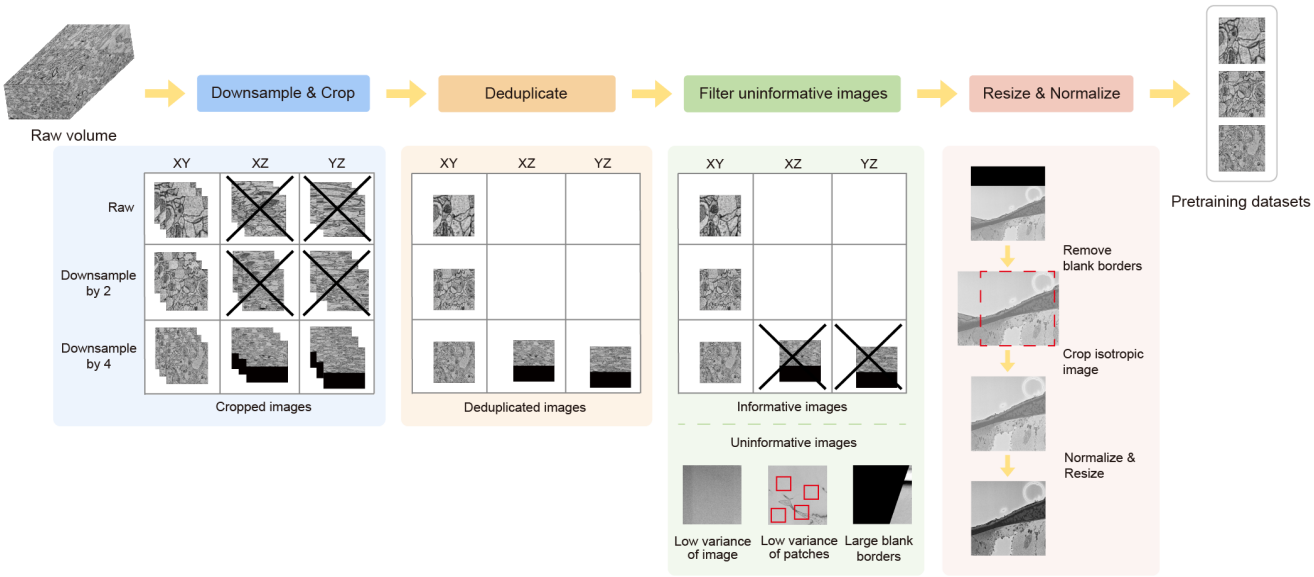
Source Data 4: values of metrics in Extended figure 4

Source Data 5: values of metrics in Extended figure 6

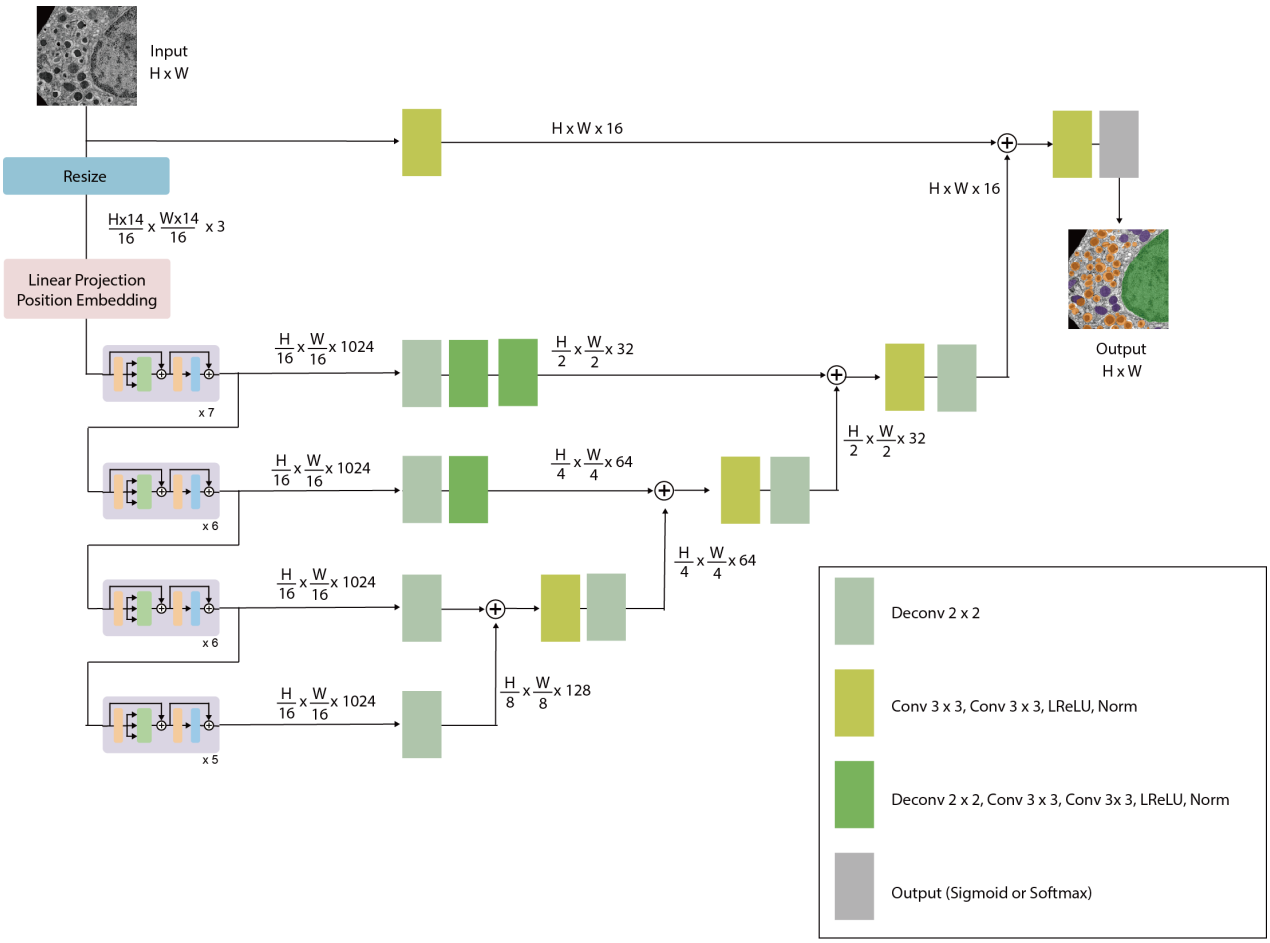
Source Data 6: values of metrics in Extended figure 7

Source Data 7: values of metrics in Extended figure 8

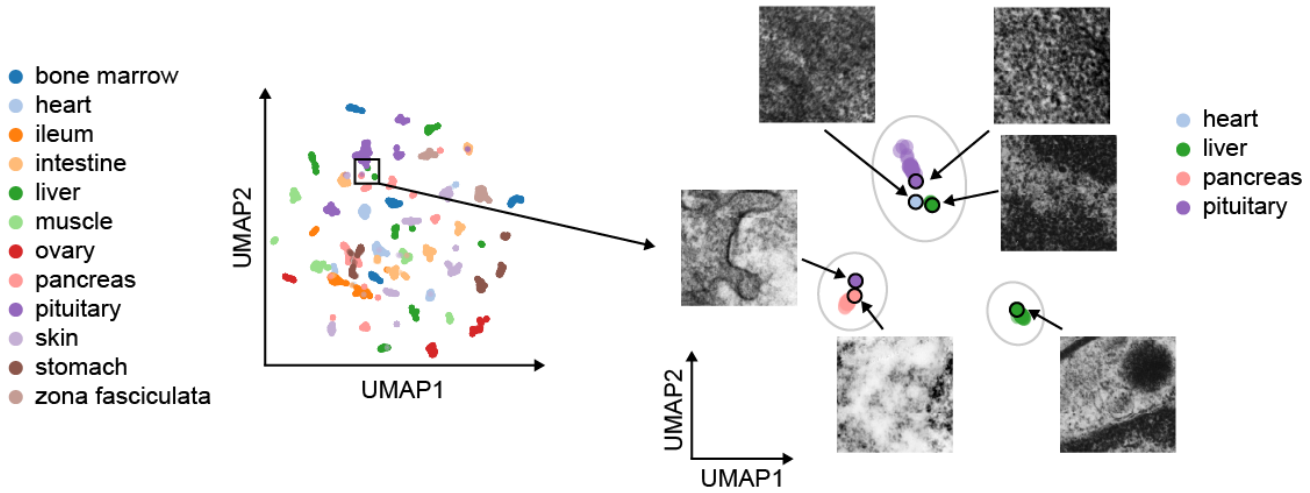
Extended Data Table 1: Sources of datasets



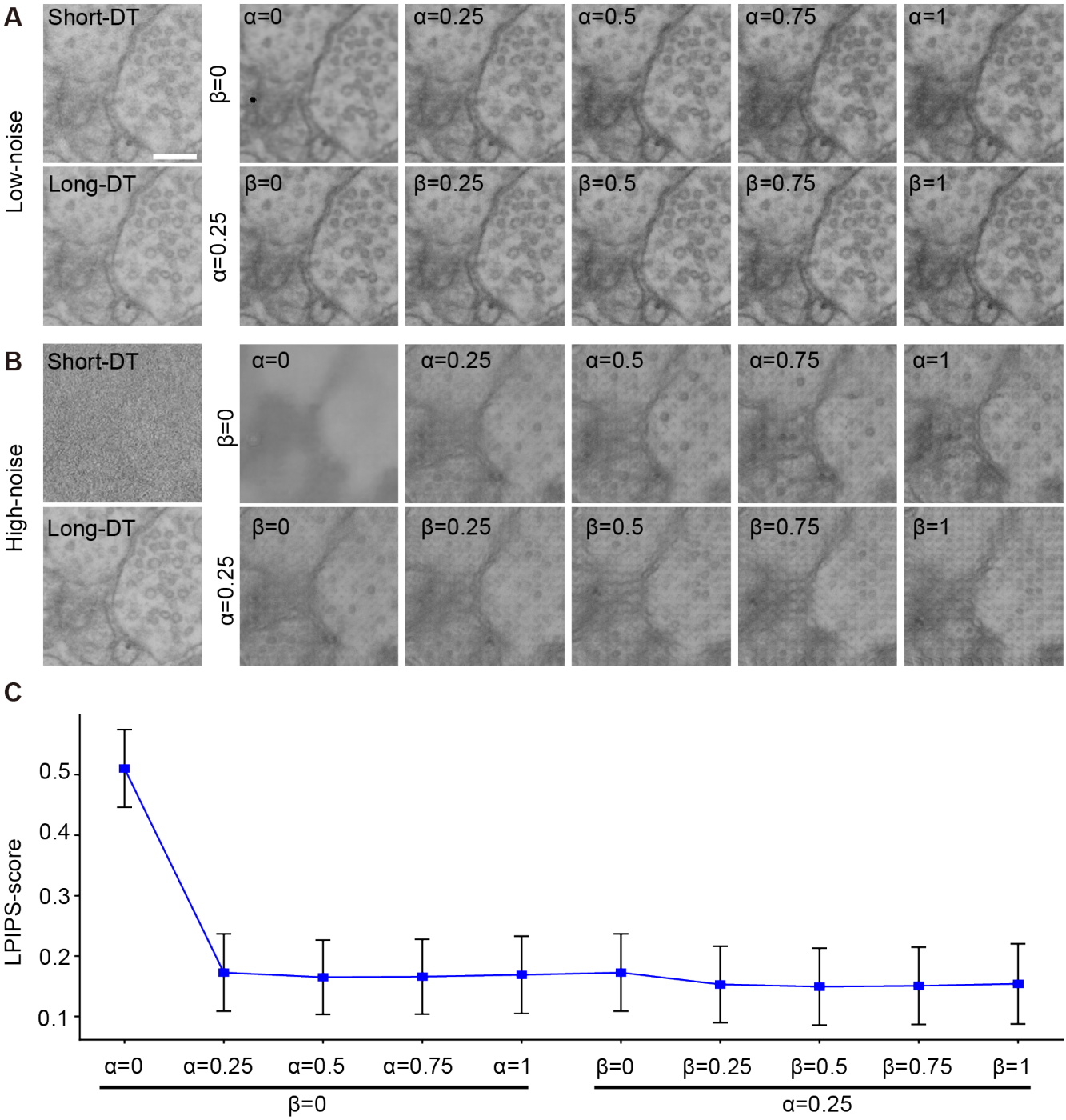
Extended Data Fig. 1. EM-5M Data Curation Pipeline. Volumetric datasets are processed through the following steps: standardization, downsampling, cropping, deduplication, uninformative region filtering, and patch-level resizing and normalization. For 2D datasets, all steps are applied except for deduplication.



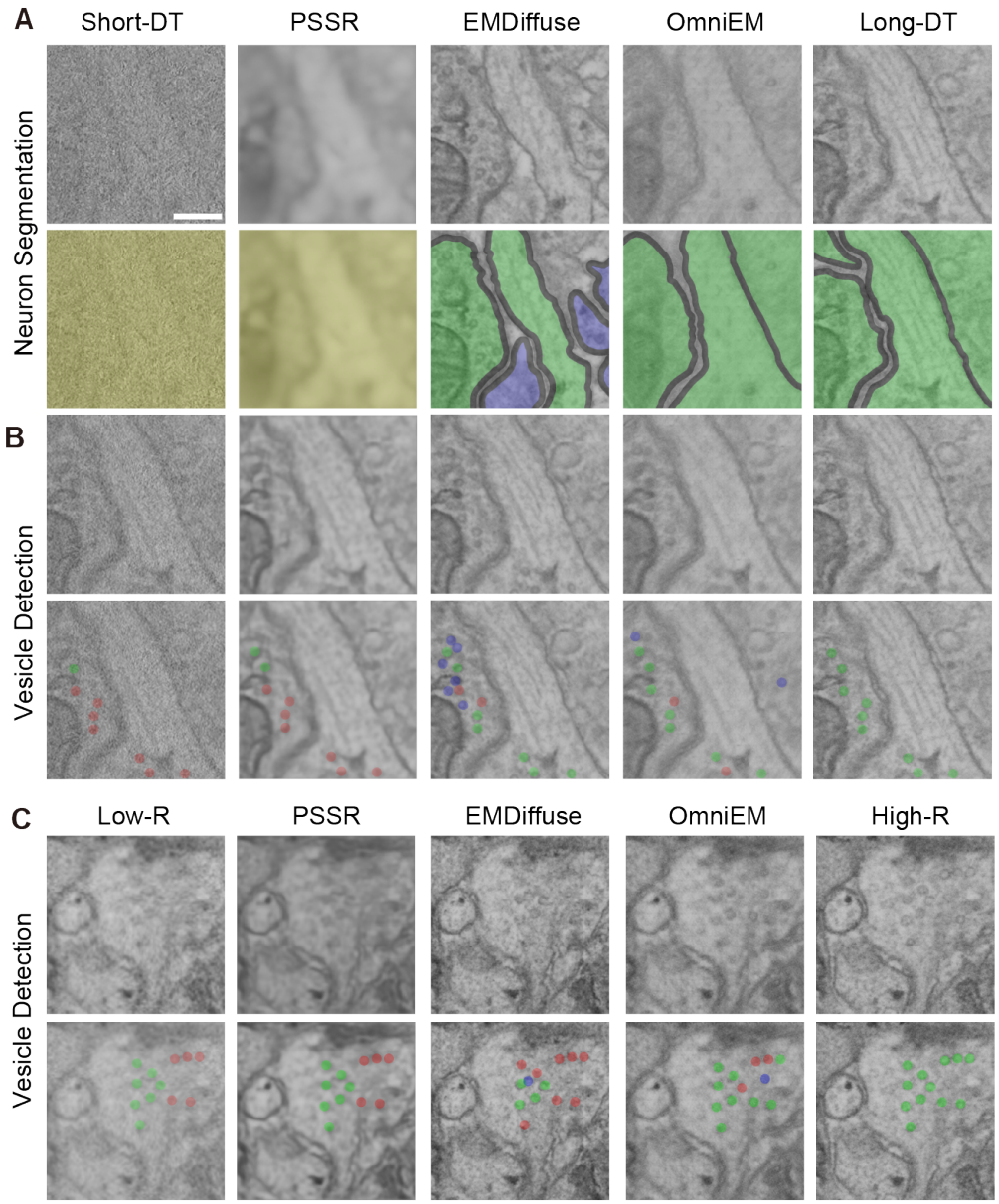
Extended Data Fig. 2. OmniEM Architecture



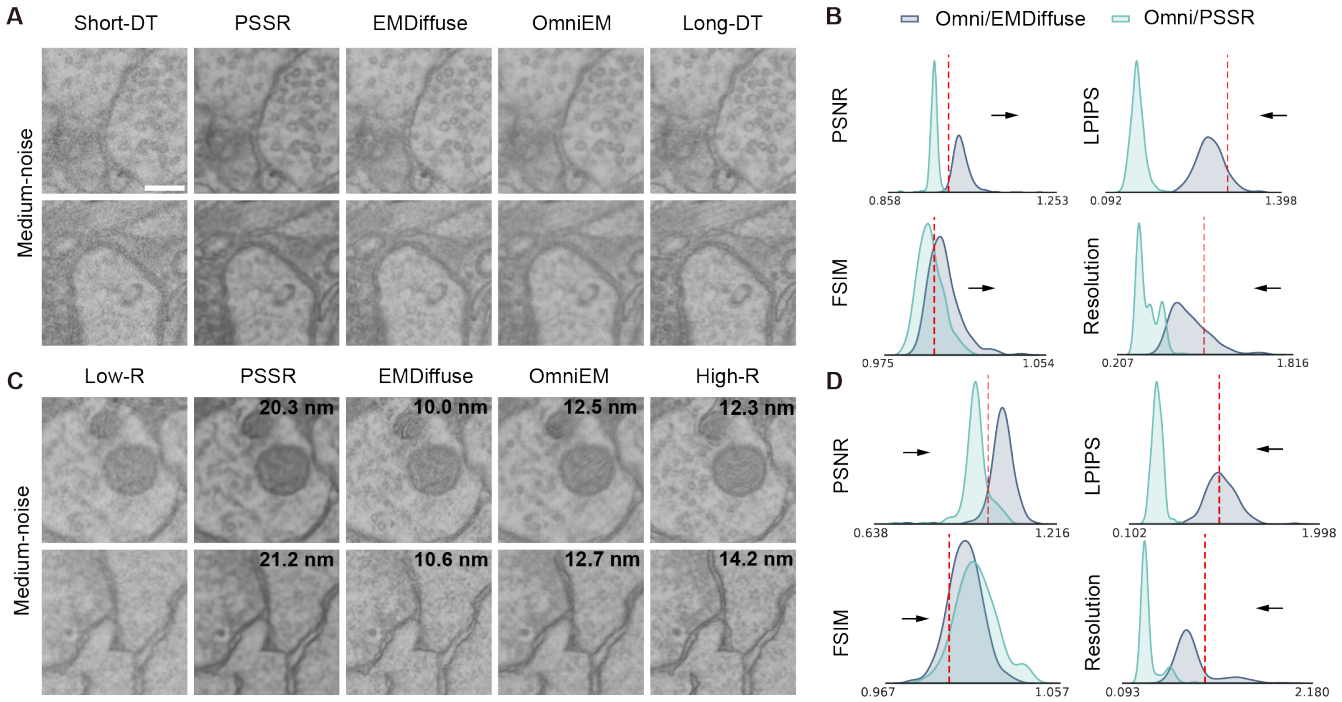
Extended Data Fig. 3. EM-DINO Distinguishes Organ Sources by Visual Features. (Left) Zoomed-in view of UMAP embedding showing images from different organs (from Fig. 2E). (Right) Representative images colored by organ cluster, illustrating how similar ultrastructural features group together.



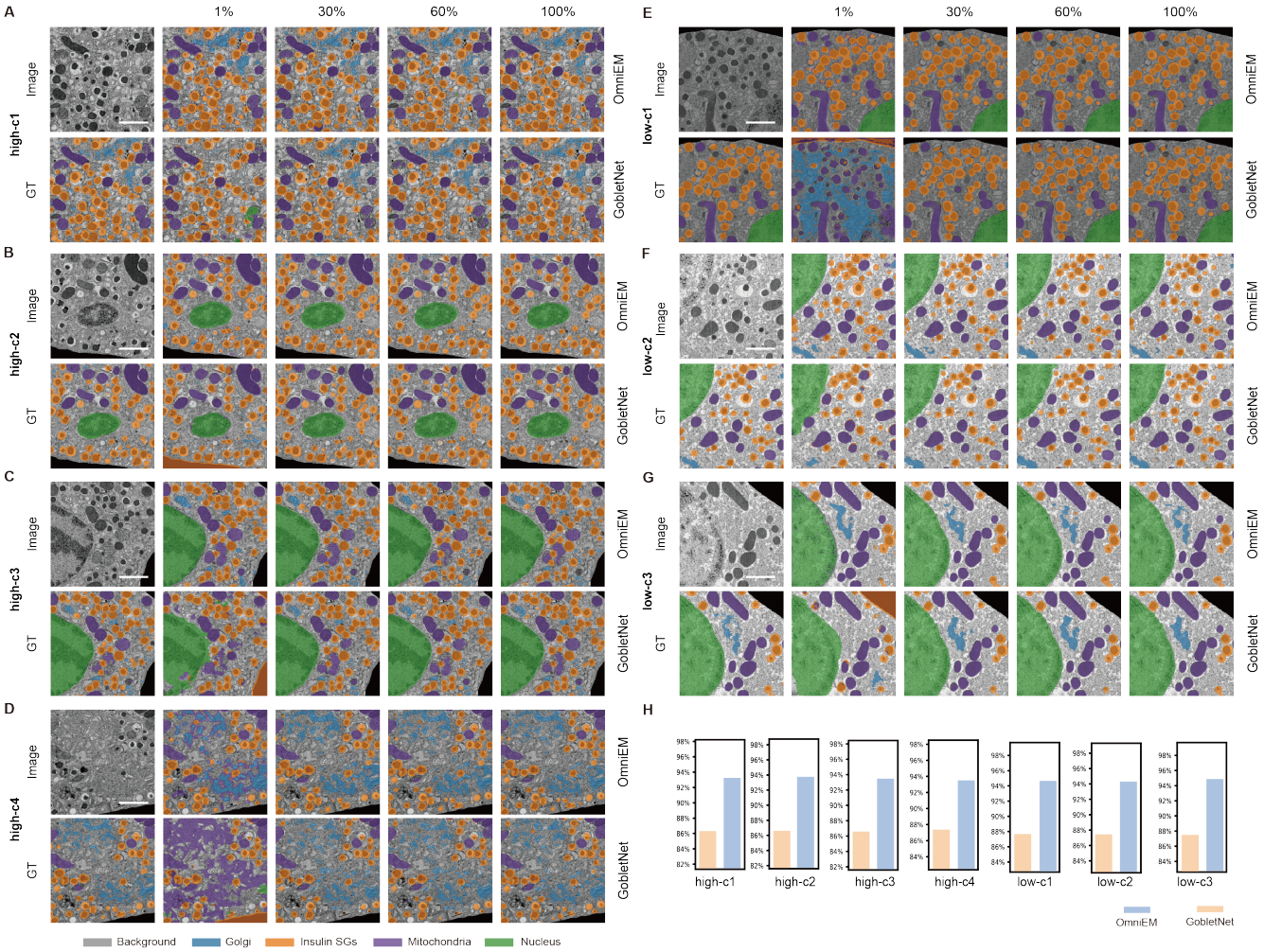
Extended Data Fig. 4. Effect of Image-Level and Feature-Level Embedding Weights on Denoising Performance. (A) Representative low-noise examples showing Short-DT inputs (leftmost column), Long-DT ground truth (bottom left), and denoised outputs as embedding weights vary. Rows correspond to $\alpha = 0.25$ (top) and $\beta = 0$ (bottom). Columns show $\beta = 0$ with α increasing from 0 to 1 (top row) and $\alpha = 0.25$ with β increasing from 0 to 1 (bottom row). As α increases, oversmoothing decreases and ultrastructural detail is progressively preserved. **(B)** Representative high-noise examples following the same layout as (A). With $\alpha = 0$, PSSR-like oversmoothing appears; as α increases, the embedding loss restores membrane and vesicle structures more faithfully. Increasing β introduces grid-like patterns that are not penalized by LPIPS. **(C)** LPIPS scores for denoised high-noise images across the range of α and β values. Introducing an image-level embedding loss ($\alpha = 0.25$) reduces LPIPS from 0.52 to 0.18; further increases in α or inclusion of β yield diminishing returns and can introduce spurious patterns. Scale bar: 200 nm.



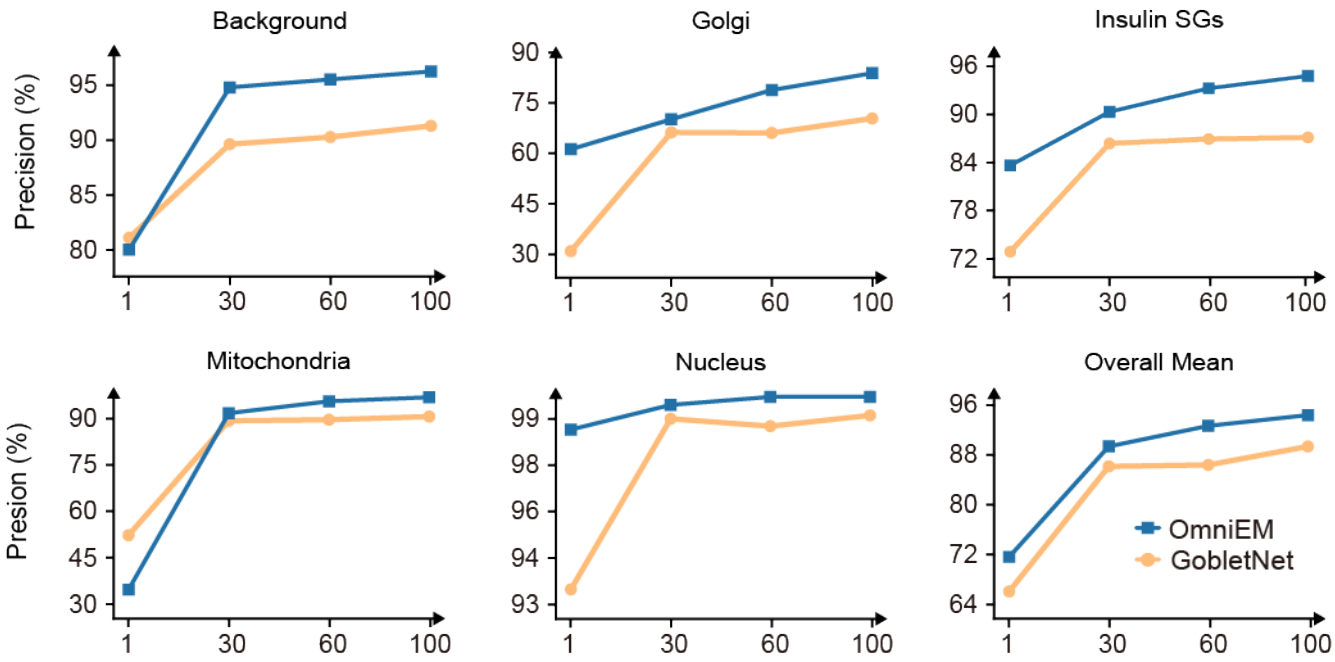
Extended Data Fig. 5. Neuron segmentation and vesicle detection comparison on different restoration results. (A) Neuron segmentation on denoised images. Top: input (Short-DT), restorations (PSSR, EMDiffuse, OmniEM), and reference (Long-DT). Bottom: manual annotations. Yellow: unannotatable regions; green: consistent with reference; blue: false positives. **(B)** Vesicle detection under the same settings. Red: missed vesicles; green: correct detections; blue: false positives. **(C)** Vesicle detection on super-resolution results. Same color coding as in (B). Scale bar: 200 nm.



Extended Data Fig. 6. Medium-level denoising and super-resolution results. (A) Comparison of the denoising results from different models under middle noise conditions. From left to right: the noisy image acquired with short dwell time (used as model input), the restored results from PSSR, EMDiffuse, and OmniEM, and the high-quality image acquired with long dwell time (used as ground truth reference). Two image regions are shown as representative cases. **(B)** Shown are the relative performance of different models on all pair-wise test data under middle noise levels, using four metrics: PSNR, LPIPS, FSIM, and resolution. Light blue line represents the ratio distribution of OmniEM to EMDiffuse, and light green line represents the ratio distribution of OmniEM to PSSR. The vertical red dashed line indicates the ratio value of 1, and arrows denote the direction of better performance for each metric. **(C)** Comparison of the super-resolution results from different models under middle noise conditions. From left to right: the low-resolution input image, super-resolved outputs from PSSR, EMDiffuse, and OmniEM, and the high-resolution ground truth image. Two representative regions are shown. The resolution of each restored image is displayed in the upper right corner. **(D)** Shown are the relative performance of different models on all pairwise test data under middle noise conditions for super-resolution, using the same four metrics.Scale bar: 200 nm.



Extended Data Fig. 7. Multi-Organelle Segmentation Results on Different Volumes from Betaseg. (A–G) Legends are the same as in Figure 4C. (H) Corresponding Dice scores of OmniEM and GobletNet trained with 100% of the training data on the cells shown in A–G.



Extended Data Fig. 8. Precision Metrics for Multi-Organelle Segmentation. The y-axis shows the precision metric for different semantic categories. Other legends are the same as in Figure 4D.