Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

EMCFsys: A scalable ecosystem for high-throughput nanoscale ultrastructure analysis across biological kingdoms

Xuping Feng

fengxp@zju.edu.cn

Zhejiang University https://orcid.org/0000-0001-9575-6916

Zeyu Yu

Zhejiang University

Jiansheng Guo

Zhejiang University

Du Mengze

Feng Liu

Nanjing Agricultural University https://orcid.org/0000-0003-1138-9472

Shan Xu

Zhejiang University

Guowei Zhang

Zhejiang University

Li Xie

Zhejiang University

Bo Han

Hong Kong Baptist University

Zhonghua Chen

The University of Adelaide

Gaoliang Deng

Hangzhou Chengfengerlai Digital Technology Co., Ltd

Chen Rui

Hangzhou Chengfengerlai Digital Technology Co., Ltd

Yong He

Zhejiang University

Article

Keywords:

Posted Date: March 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8942783/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

EMCFsys: A scalable ecosystem for high-throughput nanoscale ultrastructure analysis across biological kingdoms

Zeyu Yu[#1], Jiansheng Guo[#2], Mengze Du[#1], Feng Liu[#3], Shan Xu[1], Guowei Zhang[1], Li Xie[4], Bo Han[5], Zhong-Hua Chen[6], Gaoliang Deng[7], Chen Rui[7], Yong He[1*], Xuping Feng[*1,8]

[1]College of Biosystems Engineering and Food Science, Zhejiang University, Hangzhou, China.

[2]Center of Cryo-Electron Microscopy, Zhejiang University School of Medicine, Hangzhou, China.

[3]College of Life Sciences, Biological experiment teaching center, Nanjing Agricultural University, Nanjing, China.

[4]Analysis Center of Agrobiology and Environmental Sciences, Zhejiang University.

[5]Department of Computer Science, Hong Kong Baptist University, Hong Kong, China.

[6]School of Agriculture, Food & Wine, Waite Research Institute, The University of Adelaide, Glen Osmond, 5064 SA, Australia.

[7]Hangzhou Chengfengerlai Digital Technology Co., Ltd, Hangzhou, China.

[8]The Rural Development Academy &Agricultural Experiment Station, Zhejiang University, Hangzhou, China.

[#]These authors contributed equally to this work.

*Corresponding author

Yong He (yhe@zju.edu.cn) and Xuping Feng (fengxp@zju.edu.cn)

Abstract

Volume electron microscopy (vEM) provides nanoscale 3D visualization of cellular ultrastructure, but large-volume imaging and analysis remain limited by long acquisition times and annotation-intensive processing. Here we present the Electron Microscopy Cell Foundation ecosystem (EMCFsys), comprising a curated dataset of 4,002,802 high-quality electron micrographs spanning 14 modalities and major biological kingdoms, a foundation image restoration model (EMCellFiner) that enables 16-fold faster acquisition through high-fidelity super-resolution from low-resolution or low-dwell-time scans, and a foundation model (EMCellFound) that supports zero- to few-shot organelle classification, 2D segmentation, and precise 3D reconstruction. This facilitates high-precision 3D ultrastructure reconstruction, even for multi-piece organelles like the endoplasmic reticulum by using as little as 0.01% labeled data. Integrated into a zero-code Napari plugin, the ecosystem enables automated, high-throughput nanoscale analysis without programming expertise. By substantially reducing imaging time, annotation burden, and computational barriers, EMCF establishes an efficient, generalizable platform for large-scale vEM studies across cell types and species.

Introduction

Compared with cultured cells, *in situ* nanoscale 3D ultrastructural analysis of cells, tissues, and even organs within biological specimens are essential for a profound understanding of biological function, particularly for elucidating mechanisms underlying clinical diagnosis and pathological detection[1,2]. For instance, the 3D architecture of the endoplasmic reticulum (ER) is highly dynamic, with its tubular and sheet-like domains undergoing adaptive remodeling that is intimately linked to cellular homeostasis and disease progression[3,4].

As an emerging tool in biology[5], volume electron microscopy (vEM) has enabled breakthrough discoveries across biological systems[6–8]. This technique stands out as a pioneering approach uniquely capable of achieving computational reconstruction of nanoscale ultrastructure across substantial biological volumes, including whole

brains[9,10]. The global vEM research community is striving to leverage this technology to visualize, at nanometer resolution, the complex ultrastructure of organelles, cells, and tissues at the micrometer scale, thereby elucidating their intricate interactions with the surrounding micro-environment and enabling the construction of large-scale 3D atlases of biological specimens. However, imaging large tissue volumes remains extremely time-consuming. For example, 500mm^3 mouse brain volume would require an estimated 8.56 years for one GridTape microscope[1]. Furthermore, the lack of efficient, automated image processing pipelines continues to be a major bottleneck to the broad application of vEM in life sciences[2].

Among these limitations, image processing has emerged as a greater obstacle than image acquisition[1,2,5]. Electron microscopy data exhibits high dimensionality, complex textures, noise, low contrast, artifacts, and factors that complicate information extraction. Biological structures also vary substantially across species, cell types and physiological states, hindering generalizable models. Although deep learning-based automated analysis methods have made substantial advances in specific tasks, such as image enhancement, super-resolution, organelle segmentation[2,11–14], current supervised approaches rely on large-scale manual annotations, yielding task-specific models with limited applicability to new samples[15]. Researchers are therefore frequently forced into tedious trade-offs and combinations of manual, semi-automated, and fully automated methods, drastically constraining the efficiency of vEM-driven discovery.

Pretrained models, by leveraging large scale and diverse training data, learning generalizable feature representations that achieve exceptional performance and generalization across a wide range of downstream tasks[16–19]. The key limitation in applying this approach to electron microscopy remains the availability of suitable large-scale training data. Here, we introduce the Electron Microscopy Cell Foundation ecosystem (EMCFsys) (Fig.1) comprising three integrated components: a curated dataset of 4,002,802 high-quality electron micrographs across 14 modalities and major biological kingdoms (the electron microscopy cell foundation dataset, EMCF dataset); a restoration foundation model (EMCellFiner); and a versatile foundation model (EMCellFound) that supports zero- to few-shot organelle classification, 2D

segmentation and 3D reconstruction. For image restoration, EMCellFiner demonstrates robust zero-shot super-resolution capabilities across seven classes of image quality degradation. By enabling high-fidelity reconstruction from low-resolution or low-dwell-time acquisitions, it facilitates approximately 16-fold faster imaging while preserving subcellular detail and reducing beam-induced damage. In segmentation tasks, EMCellFound outperforms specialist models by delineation of diverse organelles and subcellular structures. It further supports high-precision 3D organelle reconstruction (for example, of the ER) using minimal labeled data (as low as 0.01% of total volume). Integrated into a zero-code Napari plugin, EMCFsys enables automated, high-throughput nanoscale analysis without programming expertise, substantially reducing imaging time, annotation burden and computational barriers for large-scale vEM studies across cell types and species.

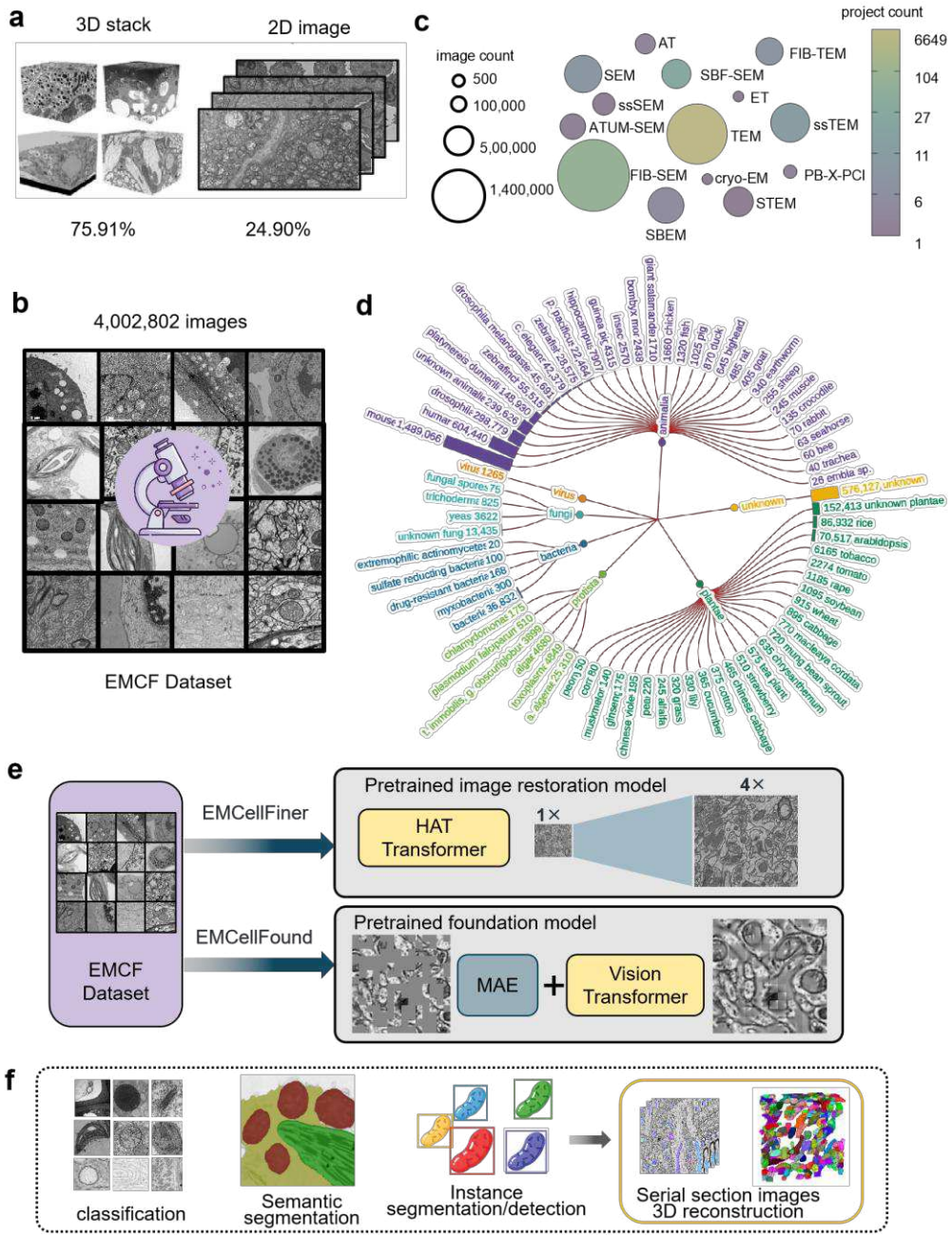


Fig. 1 Overview of Electron Microscopy Cell Foundation (EMCF) Ecosystem. a, Raw 2D and 3D electron microscopy images from diverse sources used for dataset curation. **b,** Curated EMCF dataset of 4,002,802 images spanning diverse cellular ultrastructure. **c,** Distribution of images across 14 electron microscopy acquisition methods and their respective data volume proportions. **d,** Taxonomic distribution of source specimens, spanning major biological kingdoms and species. **e,** Pretraining of EMCellFiner (image restoration) and EMCellFound (analysis foundation model) on the EMCF dataset. **f,** Adaptation of EMCellFound to downstream tasks, including zero/few-shot organelle classification, semantic segmentation, instance segmentation and 3D reconstruction.

Results

A curated cellular electron microscopy dataset

Our EMCF dataset is the first large-scale, multi-domain resource for cellular electron microscopy (**Fig. 1**a-d). Raw data consisted of 68.88% from external (public repositories, studies, and publications) and 31.12% from in-house collections. Raw images contained substantial non-informative content, including large patches lacking cellular structures after cropping, as well as severe imaging artefacts such as stripe noise, occlusion and contamination. In addition, the original data exhibited inherent heterogeneity in magnification, resolution and quality. Therefore, we implemented a systematic curation pipeline to filter, clean, and standardize the data (Supplementary Fig. 1). The final EMCF dataset comprises 4,002,802 high-quality electron micrographs spanning 14 acquisition modalities, including transmission electron microscopy (TEM), scanning electron microscopy (SEM), focused ion beam scanning electron microscopy (FIB-SEM), cryo-electron tomography (cryo-ET) and others (**Fig. 1**c). Taxonomically, the dataset encompasses five major kingdoms (Animalia, Plantae, Fungi, Protista, Monera), and Viruses (**Fig. 1**d). Images derived from Animalia constitute the largest proportion (81.29%), predominantly from key model organisms such as mouse, human, and Drosophila. Plantae images account for 8.90%, featuring model species such as *Arabidopsis thaliana* and crops (e.g., rice, tobacco). The remaining dataset fraction includes Bacteria (Monera), Fungi, Viruses, and small proportion of unclassified samples. This curated dataset, with its scale, quality and taxonomic and technical diversity, provides a robust resource for developing generalizable AI models for electron microscopy analysis in cell biology.

EMCellFiner for super-resolution

EMCellFiner is a 4× super-resolution foundation model based on the Hierarchical Attention Transformer (HAT)[20] architecture (**Fig. 2**a). In supervised evaluations with simulated degradation (5×5 Gaussian blur followed by 4-fold downsampling), EMCellFiner consistently outperformed the specialist model EMDiffuse[11], producing sharper reconstructions (**Fig. 2**c) and higher scores on Peak Signal-to-Noise Ratio (PSNR) and Feature Similarity Index (FSIM)[21] (**Fig. 2**b). The model showed robust

restoration capabilities across seven degradation types, including various blurs and noises levels (**Fig. 2**d and Supplementary Figs. 2-3). In several cases, reconstructions exceeded the sharpness of the original ground-truth images. For instance, after 3×3 Gaussian blur degradation, EMCellFiner's output was 1.29-fold sharper than the original ground truth images, whereas EMDiffuse's output was only 0.57-fold (Supplementary Fig. 2a). This advantage extended to challenging scenarios (Supplementary Figs. 4-5) and various organelles (Supplementary Figs. 6-8).

To evaluate performance in real-world settings without ground truth, we tested zero-shot generalization on diverse unseen datasets using no-reference metrics Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) and Naturalness Image Quality Evaluator (NIQE). EMCellFiner outperformed bicubic interpolation and EMDiffuse methods (**Fig. 2**e), restoring membrane clarity and cellular junctions without introducing artefacts (**Fig. 2**f). Performance remained consistent across instruments and species (Supplementary Fig. 9).

We establish EMCellFiner as a powerful, generalist pretrained model for electron microscopy image restoration. Its ability to restore high-fidelity images from lower-quality inputs has profound practical implications for experimental design. By enabling 4× super-resolution, images can be acquired images at a lower resolution (e.g., 512×512 instead of 2,048×2,048), reducing acquisition time by a factor of approximately 16-fold (Supplementary Fig. 10), for example, from 16 hours to just 1 hour, without compromising the final image quality. This acceleration dramatically increases experimental throughput and reduces operational costs.

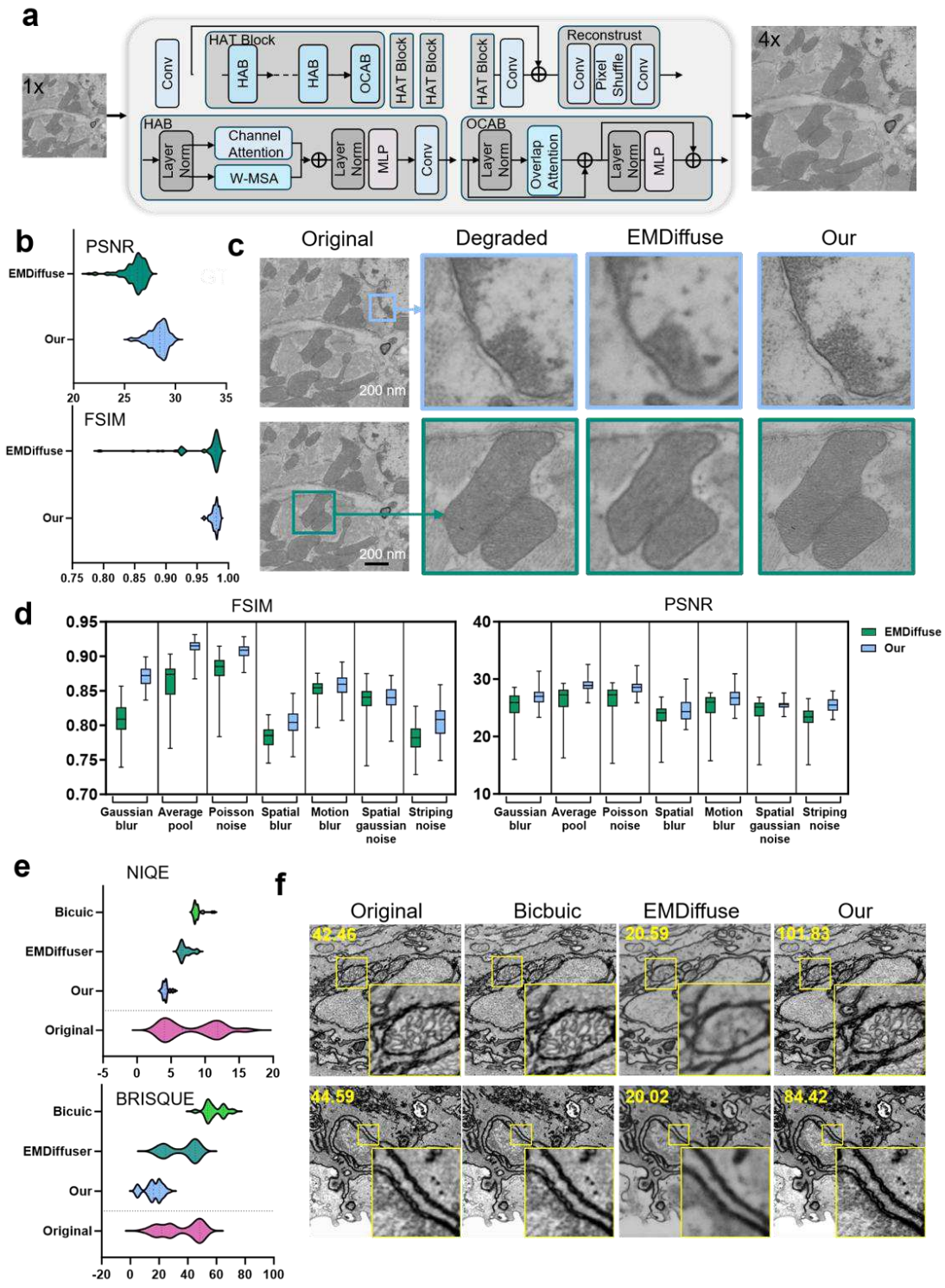


Fig. 2 EMCellFiner method and super-resolution performance evaluation. a, The architecture of EMCellFiner. Training input samples are synthetically degraded using 4x downsampling and Gaussian blur. **b,** Comparison of PSNR and FSIM for super-resolution (SR) reconstructions generated by EMCellFiner and EMDiffuse under high Signal-to-Noise Ratio conditions, specifically for images degraded by Gaussian blur. **c,** Representative SR images reconstructed by EMCellFiner and EMDiffuse. Input images for these examples were degraded by a 5×5 Gaussian blur. **d,** Comparison of PSNR and FSIM for SR reconstructions generated

by EMCellFiner and EMDiffuse across various degradation patterns. **e,** Comparison of BRISQUE and NIQE for SR reconstructions by Bicubic interpolation, EMCellFiner, and EMDiffuse under low Signal-to-Noise Ratio conditions. **f,** Representative SR images reconstructed by EMCellFiner and EMDiffuse under low Signal-to-Noise Ratio conditions. Sharpness values are shown for related images. The data sources used for these tests are detailed in Supplementary Table 1.

EMCellFound learns discriminative features for organelle classification

We also developed EMCellFound, a generalist model for cellular feature representation, by employing a Masked Autoencoder[22] (MAE) self-supervised strategy on EMCF dataset (**Fig. 3**a). To benchmark its performance, we compared MAE with leading self-supervised learning methods (DINO[23], MoCov3[24], and All4One[25]) on a challenging 8-class organelle classification task (nucleus, mitochondria, ER, Golgi apparatus, cell wall, chloroplast, lipid droplets, and endosome). With the pretrained Vision Transformer backbone frozen, a k-nearest neighbors (KNN)[26] classifier was applied to the learned features (**Fig. 3**b). Under identical training conditions, the MAE-pretrained EMCellFound achieved a mean accuracy of 96.01%, significantly outperforming all other strategies (**Fig. 3**c). Supervised classification performance and t-SNE[27] visualization of the feature space demonstrated clear separation of clusters corresponding to the eight organelle classes (**Fig. 3**d,e). Partial overlap was observed between Golgi apparatus and endosome clusters, consistent with their shared vesicular morphology and functional connection via intracellular trafficking pathways[28].

Membrane-bound organelles are enclosed by their own distinctive lipid bilayer, which delineates an internal lumen separate from the cytoplasm and confers their characteristic morphology. Attention maps confirmed that EMCellFound acquires biologically informative, discriminative features solely through self-supervised pretraining on heterogeneous electron microscopy data. MAE pretraining facilitated precise reconstruction of fine membrane ultrastructure, including sharp boundaries, despite 75% patch masking (Fig. 3e and Supplementary Figs. 11–12).

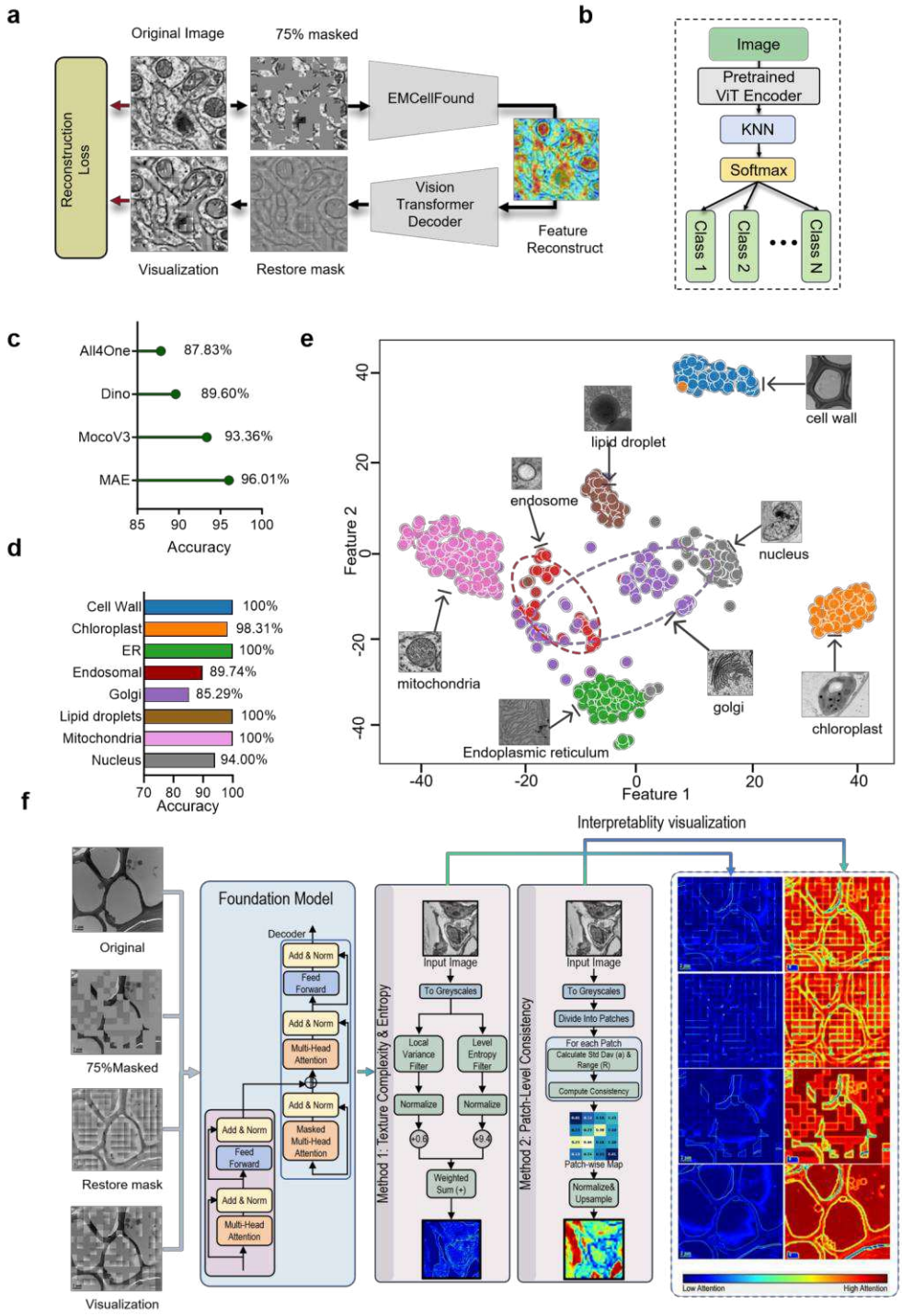


Fig. 3 EMCellFound Performance in Zero-Shot Organelle Classification. **a**, Architecture of the EMCellFound model based on the Masked Autoencoder (MAE) self-supervised learning framework. **b**, KNN classification architecture. **c**, Comparison of classification performance of ViT base models pretrained with different self-supervised algorithms (MAE, DINO, MoCo v3, All4One). **d**, Classification performance across the eight organelle classes using EMCellFound. **e**, t-SNE visualization of feature distributions for eight organelle classes, extracted from the EMCellFound encoder. **f**, Attention heatmaps generated by two analysis methods: Texture Complexity&Entropy analysis and Patch-Level Consistency analysis. Color bars represent attention scores, with colors ranging from blue (low importance) to red (high

importance), derived from the corresponding analysis metrics.

EMCellFound for semantic segmentation and 3D reconstruction

To evaluate transferability to downstream semantic segmentation tasks, we tested EMCellFound on the Plantorganelle Hunter dataset[12], which includes four organelle classes. We compared different fine-tuning strategies (Supplementary Fig. 13) against scratch-trained baselines (Unet[29], ViT[30], Mask2former[31]) and the domain-specialist model OrgSegNet[15]. Models initialized with EMCellFound weights consistently outperformed scratch-trained baselines across all settings. Full fine-tuning yielded the highest mean Intersection over Union (mIoU), exceeding that of OrgSegNet while requiring fewer epochs to converge (**Fig. 4**a,b; Supplementary Fig. 14). Organelle-specific performance was comparable between EMCellFound-based models (EMCF-Finetune and EMCF-Mask2Former) and OrgSegNet (**Fig. 4**c). For chloroplasts (abundant class), all models reached mIoU > 90%. For the less frequent and imbalanced nucleus class, full fine-tuning outperformed the specialist model.

EMCellFound-based models achieved peak performance with substantially less training data: EMCF-Finetune reached optimal mIoU using only 10% of the dataset, compared with 100% required for scratch-trained ViT and 60% for OrgSegNet under matched conditions (**Fig. 4**d). This indicates effective transfer of learned structural priors, significantly reducing the reliance on large annotated datasets.

We next applied EMCellFound to 3D reconstruction from sparse 2D annotations, an approach that requires fewer GPU resources than direct 3D segmentation. On the Liver-6[32] vEM dataset, fine-tuning with annotations from only 0.01% of the total volume produced accurate 3D models of mitochondria, nuclei and lipid droplets, outperforming a standard 3D U-Net (**Fig. 4**e, f). The improvement was particularly pronounced for ER, which comprised complex, non-enclosed tubular and sheet-like domains. EMCellFound generated topologically continuous and smooth surfaces, preserving membrane network integrity, as reflected in higher volume and surface area metrics compared with 3D U-Net (**Fig. 4**f; Supplementary Fig. 15).

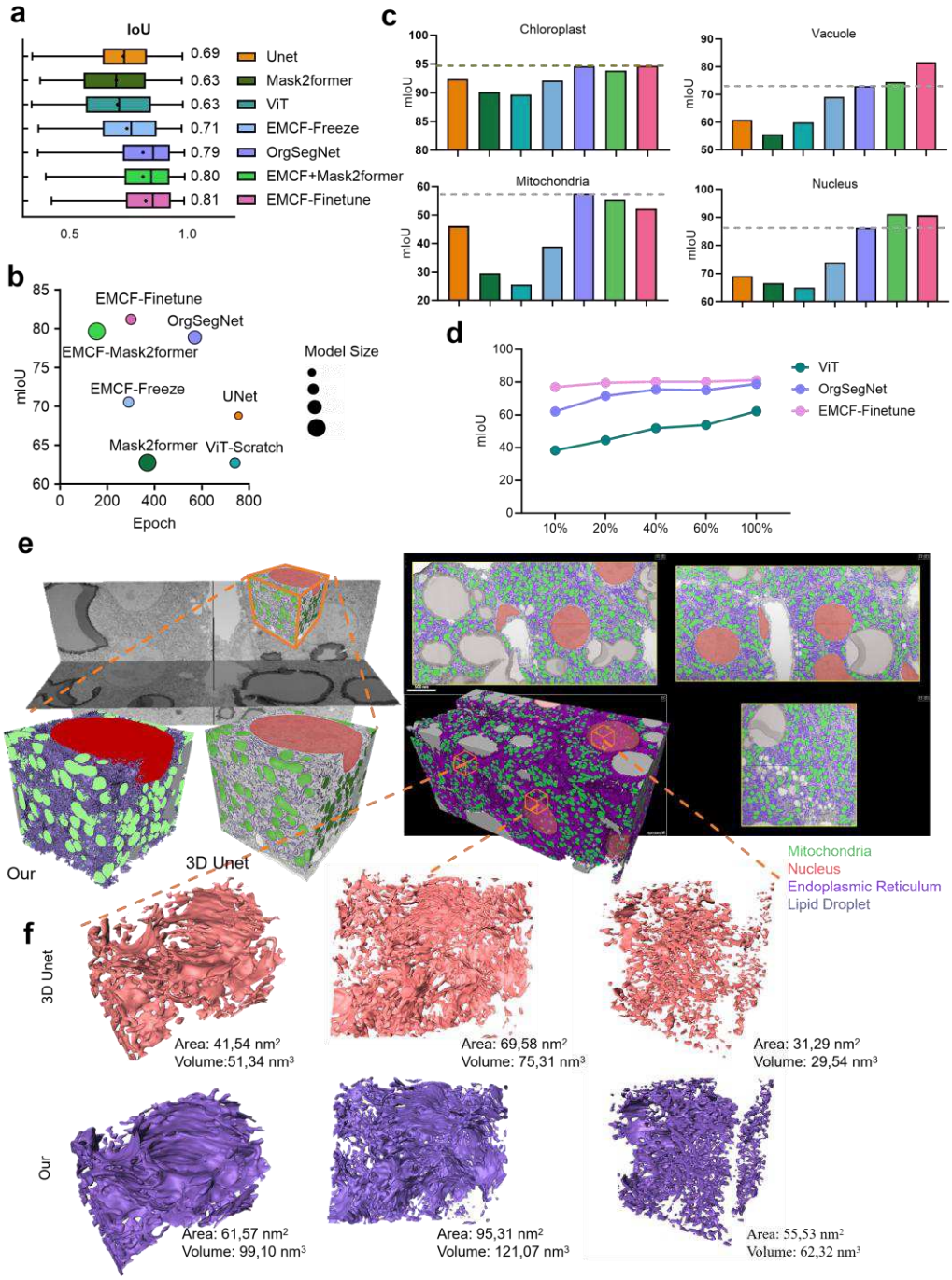


Fig. 4 EMCellFound performance in semantic segmentation task. a, Semantic segmentation results of four baseline models (Unet, ViT, Mask2Former, OrgSegNet) and three fine-tuned EMCF (EMCF-Freeze, EMCF-Mask2Former, and EMCF-Finetune) on the Plantorganelle Hunter dataset. **b**, Comparison of segmentation performance across seven models based on the mIoU, Epoch and model size in the inference process. **c**, Organelle-specific segmentation performance across seven models. **d**, Segmentation results of ViT, OrgSegNet, and EMCF-Finetune with increasing training dataset size. **e**, Representative 3D reconstruction of organelles from Liver-6 vEM data using EMCF-Finetune. A selected volumetric region from the Liver-6 vEM dataset is segmented in 2D slices by EMCF-Finetune and subsequently reconstructed into a 3D model. Lipid droplets (LD), endoplasmic reticulum (ER), mitochondria,

and nuclei are shown in gray, purple, green, and translucent red, respectively. **f**, Comparative 3D reconstruction of the intricate ER network from **e** using a baseline 3D U-Net (top, red) and EMCF-Finetune (bottom, purple). Quantitative metrics (total surface area and volume) highlight the improved topological completeness and smoothness achieved by EMCF-Finetune, overcoming fragmentation issues seen in the baseline.

Note: **a,** Unet, Mask2former (ViT as backbone with Mask2former as segmentation head), ViT (ViT as backbone with UperNet[33] as segmentation head) are scratch-trained models. EMCF-Freeze is the model with the backbone weight freezing and only trains the UperNet segmentation head. EMCF-MaskFormer is the model consist of EMCellFound backbone with Mask2former segmentation head, and EMCF-Finetune is the model with EMCellFound backbone with Uperhead segmentation head. see Supplementary Fig. 13.

Few-shot instance segmentation with EMCellFound.

To evaluate generalization and data efficiency for instance segmentation, we tested EMCellFound on three public electron microscopy datasets of mitochondria: Bock[34], Kidney[35], and Lee[36] (**Fig. 5**). We systematically compared models initialized with EMCellFound weights against equivalent architectures trained from scratch, encompassing both standard CNN backbone (RTM[37]) and Transformer-based (ViT and ViTAdapt[38]) backbones (model configurations are in Supplementary Fig. 16). Scratch-trained RTM underperformed across segmentation and detection tasks. Scratch-trained ViT-RTM and ViTAdapt-RTM also yielded lower mean Average Precision (mAP) for instance segmentation on all three datasets compared with models using pretrained EMCellFound weights (**Fig. 5**a), consistent with visual results (**Fig. 5**b). Incorporation of pretrained weights in EMCF-Finetune and EMCFAdapt-Finetune improved mAP for both small and large mitochondria (**Fig. 5**c).

We next compared EMCellFound to MitoNet[14], a specialist model for mitochondrial instance segmentation trained on 21,860 serial sections. Fine-tuning EMCellFound on only 30 images (1,024 × 1,024 pixels) from the Kidney dataset yielded performance comparable to MitoNet (**Fig. 5**d and 5e). Our findings underscore a transformative shift in paradigm that EMCellFound eliminates the need for prohibitive large-scale dataset curation, delivering high-precision image analysis with minimal supervision.

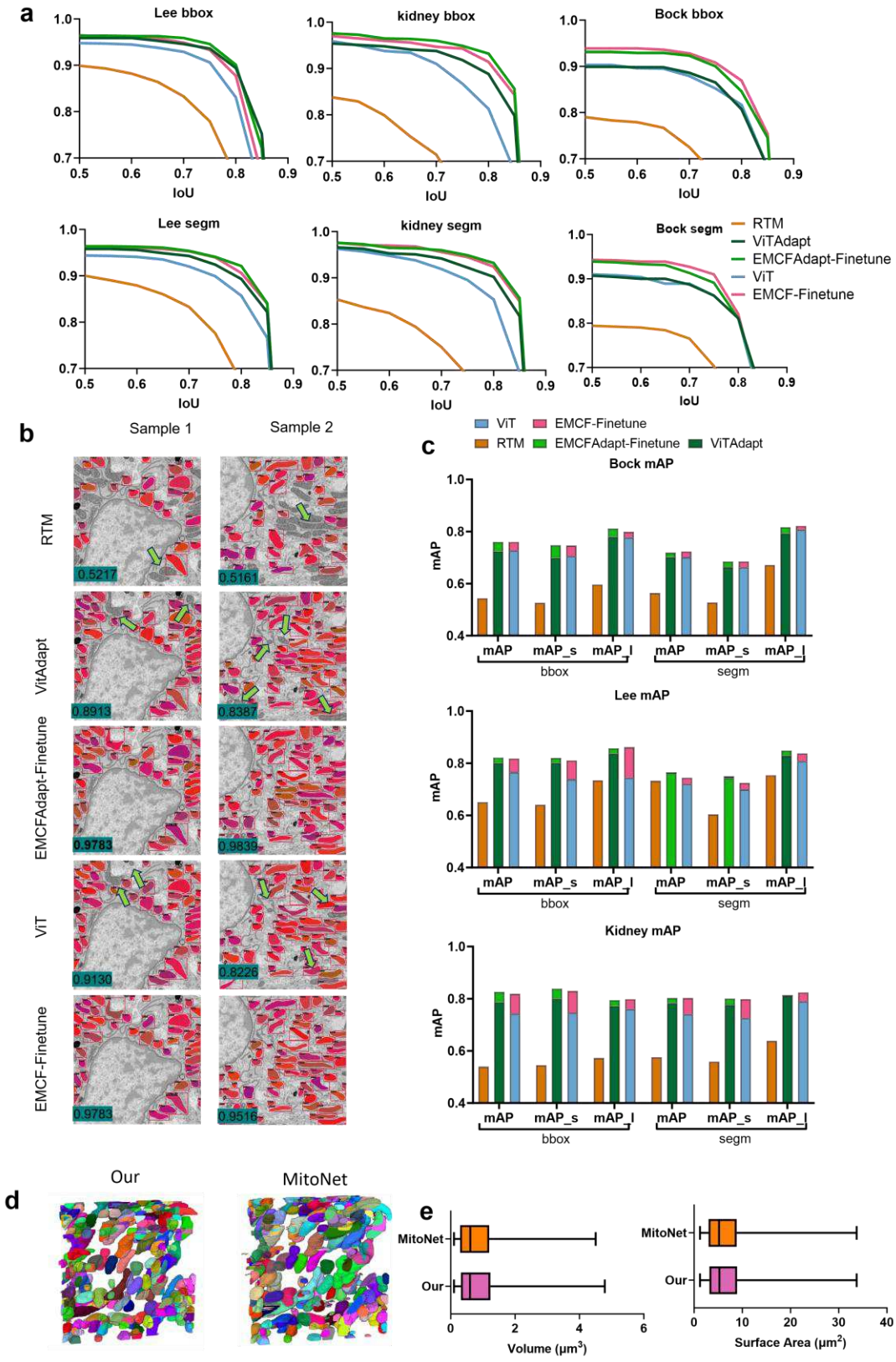


Fig. 5 Application of EMCellFound for instance segmentation of mitochondria. a, Mitochondrial instance segmentation results of five models on Bock, Lee, and kidney datasets. The first and second columns represent the mean Average Precision (mAP) for bounding box

detection and segmentation, respectively. The strategies of the five models are detailed in Supplementary Fig. 17. **b**, Representative instance segmentation of mitochondria by five models on kidney data. The numbers within the images represent the bounding box recall (Recall = correctly detected instances / total ground truth instances). **c,** Detection and instance segmentation performance of models across different object sizes. mAP denote the mean Average Precision, mAP_s for small objects (bounding box size < 96×96 pixels), and mAP_l for large objects (bounding box size > 96×96 pixels), respectively. **d,** 3D Mitochondrial reconstruction from Kidney data following automatic segmentation by EMCellFound and MitoNet. **e,** Relations of surface area and volume of mitochondria obtained from EMCellFound and MitoNet.

EMCFsys enables efficient 3D reconstruction for both time imaging saving and accuracy.

High-resolution volumetric imaging of large biological specimens usually encounters a prohibitive trade-off between acquisition speed and structural fidelity (details or accuracy). While vEM techniques like FIB-SEM offers nanoscale precision, acquiring dense volumes is notoriously time-consuming. EMCellFiner debunks this trade-off by computationally upscaling the rapid low-resolution scan, to reconstruct a high-fidelity 4 times upsacle volume in 6 hours of offline computation using a RTX3090 GPU (**Fig. 6**a). The time savings scale with volume size: larger datasets benefit from low-resolution/low-dwell-time acquisition followed by fast reconstruction. Applying EMCellFiner to unseen volumes of anterior pituitary and intestinal epithelial cells yielded images with improved delineation of organelle membranes, thereby enhancing the accuracy of subsequent segmentation tasks (**Fig. 6**b-c).

To demonstrate the synergy between restoration model EMCellFiner and analysis foundation model EMCellFound, we employed a few-shot learning strategy, fine-tuning EMCellFound on just 3 randomly selected slices to segment the entire volume in both tissues. EMCellFiner preprocessing significantly boosted segmentation accuracy, with the greatest gains observed for the ER (Supplementary Fig. 17). Following combined restoration and segmentation, the detected ER-mitochondria and ER-secretory granules contact areas were substantially enlarged, enabling more accurate characterization of membrane appositions and their potential functional roles (**Fig. 6**g). EMCFsys supports automated segmentation not only of closed, compact structures in 2D planes (e.g., nuclei, secretory granules, mitochondria) but also

accelerates reliable identification and annotation of irregularly shaped, fine-grained, or linear features, including plasma membranes in pollen tubes (Supplementary Video 2), cell membranes in human anterior pituitary cells, and the intricate reticular network of the ER (Supplementary Fig. 18). Together, these tools break down barriers to large-scale volume electron microscopy by enabling efficient high-quality data generation and minimal-annotation 3D reconstruction, thus accelerating accurate nanoscale biological discovery across diverse organisms.

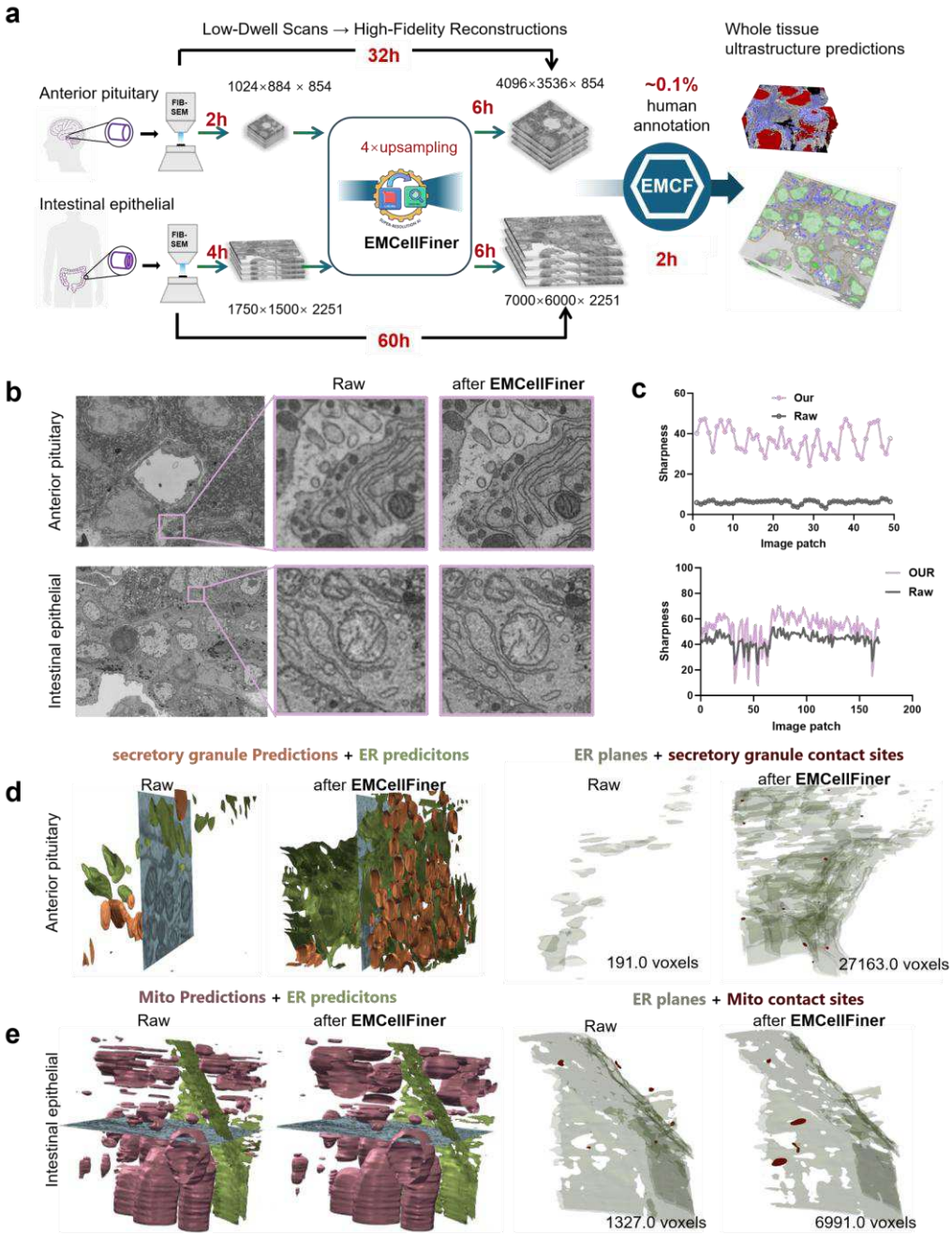


Fig. 6 Performance of EMCFsys pipeline for 3D reconstruction of human cells and tissues. **a**, Workflow illustrating substantial reduction in total acquisition time and manual annotation effort through the combination of super-resolution restoration by EMCellFiner and automated annotation by EMCellFound. **b,** zoomed-in representative SR images of anterior pituitary cells and intestinal epithelial cells processed by EMCellFiner. **c**, Quantitative comparison of image sharpness scores across patches from large-scale volumes processed with and without EMCellFiner. **d**, Comparison of 3D cellular ultrastructure reconstructions from original data versus EMCellFiner-enhanced super-resolution data of anterior pituitary cells. Left panels show a 3D rendering of the ER predictions and secretory granules predictions, and the right panels show the contact site regions (red) where ER and secretory granules are within each other. **e**, Comparison of 3D cellular ultrastructure reconstructions from the original data versus EMCellFiner-enhanced super-resolution data of

intestinal epithelial cells. Left panels show a 3D rendering of the ER predictions and mitochondria (Mito) predictions, and the right panels show the contact site regions (red) where ER and mitochondria are within each other.

Discussion

Here, we introduce EMCFsys, comprising a curated dataset of 4,002,802 high-quality images spanning multiple modalities and species, alongside two self-supervised foundation models (EMCellFiner and EMCellFound) for end-to-end electron microscopy (EM) image analysis.

Large scale tissue imaging remains extremely time-consuming, for example, acquiring a complete mouse brain volume would require an estimated 139.6 years using one EM microscope with a beam deflecting mechanism[39]. Preserving *in situ* 3D organization of cells within such samples of paramount importance for studies of tissue biology and pathological conditions. Our findings demonstrate that the synergistic integration of EMCellFiner and EMCellFound substantially reduces imaging time while requiring only minimal fine-tuning on a few images to achieve accurate organelle segmentation (**Fig. 6**). EMCFsys enables precise 3D reconstruction of irregular membrane-bound organelles, such as the ER, and reveals additional ER-secretory granule contacts in human anterior pituitary cells, as well as expanded ER-mitochondria interfaces in intestinal epithelial cells. This approach is particularly effective for complex membrane structures, including the tubular and sheet-like domains of the ER and certain plasma membranes, thereby facilitating a more integrated understanding of subcellular functions and intercellular interactions within tissues.

Leveraging the HAT architecture, EMCellFiner delivers state-of-the-art super-resolution performance, surpassing existing models in reconstructing intricate cellular ultrastructure across scales and diverse data sources from various EM techniques. Critically, for large biological specimens and datasets, EMCellFiner's capacity to recover fine details from low-resolution or low-dwell-time acquisitions minimizes imaging time and costs while preserving or enhancing image quality. This breakthrough paves the way for more efficient experimental designs to tackle key research questions, especially in large volume vEM studies.

In parallel, EMCellFound, pretrained using masked autoencoding (MAE), captures biologically meaningful features comparable to those discerned by experienced cell biologists (**Fig. 3**). It exhibits state-of-the-art transferability in organelle recognition, detection, semantic segmentation, and instance segmentation tasks, often matching or exceeding specialized models with just a fraction of labeled data. Moreover, it supports 3D reconstruction from sparse 2D annotations, yielding improved topological continuity for non-enclosed ultrastructure (**Fig. 6**), such as irregular ER membranes and plasma membranes.

We have demonstrated that EMCellFiner and EMCellFound trained on vast EM datasets deliver substantial improvements in downstream task performance. However, their computational demands, particularly the reliance on large-scale datasets and high computing resource requirements, remain a practical limitation. Future optimizations focusing on few-shot learning, lightweight architecture, or employing token-based reconstruction objectives may mitigate these constraints. Furthermore, integrating large language models and generative models offer promising avenues to establish structured memory for continual learning, and incorporating scientific knowledge, biological principles, and theories through machine learning tools for feature structuring, further enhancing model adaptability and interpretability.

Finally, our models, dataset and Napari plugin are publicly released to support adoption and extension by the community. We anticipate that this open resource will facilitate automated, high-throughput analysis of vEM data across diverse biological systems.

Methods

Preparation of the Electron Microscopy Cell Foundation Dataset

To establish a comprehensive electron microscopy dataset for large-scale cellular representation learning, we systematically collected and curated a vast corpus of electron microscopy data (**Fig. 1**). The dataset integrates multi-source electron microscopy images from diverse origins, including public repositories (EMPIAR, BossDB, OpenOrganelle, and Nanotomy), datasets accompanying peer-reviewed

publications, and shared archives from collaborating laboratories across China, the United States, Japan, and the Netherlands.

The collected data encompasses a broad spectrum of biological taxa, spanning animal, plant, and microbial specimens, and include both 2D and 3D imaging modalities. Multiple electron microscopy platforms were employed, such as TEM, SEM, FIB-TEM, FIB-SEM etc. This diversity ensures comprehensive coverage of subcellular structures, including membranes, organelles, and fine ultrastructural details across a wide range of biological contexts.

EMCFD Data Cleaning, Augmentation, and Quality Control

To construct a high-quality electron microscopy foundation dataset, we designed a multi-stage data processing pipeline encompassing slicing, cropping, augmentation, redundancy removal, and automated quality control (Supplementary Fig. 1). Distinct strategies were applied to 2D and 3D electron microscopy images to ensure both data diversity and structural integrity.

For 2D electron microscopy images, cropping was performed according to their native resolutions. Small-scale images from 512×512 to 4,096×4,096 pixels were partitioned into 1, 4, or 9 partially overlapping tiles, with a 20% overlap introduced to preserve contextual continuity across adjacent patches. Images exceeding 4,096×4,096 pixels were systematically cropped into 512, 768, 1,024, and 1,536 pixels patches using a top-left to bottom-right scanning scheme. This process ensured balanced coverage of fine subcellular regions while avoiding boundary artifacts.

Given the high inter-slice redundancy inherent in vEM stacks, we adopted a sparse and multi-resolution slicing strategy. From each 3D stack, a subset of 2D slices (denoted as m frames) was uniformly sampled at fixed intervals to avoid oversampling highly correlated slices. Subsequently, we performed directional slicing along the X-Y, Y-Z, and X-Z planes to capture anisotropic structural information. Each extracted 2D slice larger than 5000×5000 pixels was cropped into image blocks of multiple scales, 512×512, 1,024×1,024, 1,536×1,536, 2,048×2,048, 3,072×3,072, and 4,096×4,096 pixels, following a left-to-right, top-to-bottom traversal pattern. To ensure non-redundancy, each original slice was processed only once per target size, and distinct

slices were used for different resolutions. This strategy allowed the dataset to retain rich multiscale information while maintaining low inter-patch redundancy.

To further enhance dataset reliability, we trained a ResNet50[40] binary classifier to distinguish high-quality EM images from those exhibiting noise, defocus, or mechanical artifacts. The classifier was trained on a manually labeled subset of 50,000 samples annotated by expert microscopists. Low-quality or corrupted patches were automatically discarded.

A large-scale redundancy filtering pipeline was implemented to eliminate highly repetitive or overlapping image patches. We adopted a perceptual hashing-based approach for efficient similarity detection. Each image was converted into a compact hash representation capturing its global visual structure and tonal distribution. Pairwise Hamming distances between hashes were then computed, and image pairs with distances below a threshold of 5 were considered visually redundant. For each redundant group, only one representative image was retained, effectively reducing data duplication while preserving dataset diversity.

After all filtering and quality control stages, a total of 4,002,802 EM images of varying resolutions from 512×512 to 4,096×4,096 pixels were retained. The resulting dataset, EMCF dataset, represents one of the largest and most structurally diverse EM image collections to date, encompassing multiscale, multi-organism, and multi-modality features suitable for large-scale foundation model pretraining.

Image Restoration Model: EMCellFiner Design and Training

To reconstruct high-quality electron microscopy images from degraded observations, we developed EMCellFiner (Electron Microscopy Foundation Finer model), a transformer-based super-resolution foundation model built upon the HAT framework. The model learns a mapping function $F: L -> H$, where L and H denote the low-quality and high-quality EM image domains, respectively. The objective is to predict a high-fidelity reconstruction $F(L) \approx H$, effectively restoring fine subcellular textures and structural integrity. To simulate realistic degradation processes in EM imaging, we designed a two-stage stochastic degradation pipeline to generate the low-quality images L from the pristine high-resolution images H (sampled from the EMCF

dataset). This degradation process aims to mimic various noise, blur, and compression artifacts commonly introduced during electron microscopy acquisition. Each high-resolution image was first subjected to random upsampling and downsampling to simulate changes in magnification and scanning area.

Stage1: A random average blur with kernel size in the range [1, 21] was then applied to imitate optical defocus and beam-induced diffusion. Subsequently, the image was downsampled using one of three randomly selected interpolation modes, including bicubic, bilinear, and area sampling, followed by the addition of Gaussian noise range 1 to 30, Poisson noise, and JPEG compression artifacts.

Stage2: A second round of random blurring with the kernel from 1 to 21 and then 4× downsampling was performed to achieve the target degradation level, followed by an additional Gaussian noise injection. This two-step degradation ensures that the generated images encapsulate a wide spectrum of realistic low-quality conditions observed in diverse EM imaging setups.

EMCellFiner Model Architecture and Training Strategy

Following the configuration of HAT, EMCellFiner adopts a hierarchical transformer architecture with hybrid attention mechanisms for spatial and channel-wise feature refinement.

We employed a GAN-based training scheme, where EMCellFiner acts as the generator, and a lightweight CNN-based discriminator (VGG-19) guides perceptual realism.

The overall objective function combines pixel-level reconstruction loss, perceptual loss, and adversarial loss:

$$L = \lambda_1 L_1 + \lambda_2 L_{perc} + \lambda_3 L_{GAN}$$

where L_1 denotes the L1 loss between the restored and ground-truth images, L_{perc} is the perceptual loss computed on VGG feature space, and L_{GAN} corresponds to a vanilla adversarial loss encouraging photorealistic reconstruction.

Training was conducted on 8 NVIDIA H800 GPUs for 400,000 iterations with a batch size of 16. The entire process took approximately 4 days. Adam optimizer was used with an initial learning rate of 2×10^{-4}, decayed following a cosine annealing

schedule. All images were randomly cropped into 64×64 patches as input, and the model was trained to reconstruct corresponding 256×256 high-resolution outputs.

EMCellFound architecture and implementation

EMCellFound is an encoder pretrained within a masked autoencoding framework consisting of a paired encoder-decoder architecture. To achieve both scalability and computational efficiency, a 12-layer Vision Transformer (ViT) encoder was adopted to model long-range dependencies and capture fine structural representations typical of EM imagery. A lightweight 9-layer Transformer served as the decoder, with an embedding dimension of 512, responsible for reconstructing the masked portions of the input image. This asymmetric encoder–decoder design balances representation capacity and training efficiency, allowing EMCellFound to learn high-level semantic priors while maintaining feasible training costs for large-scale EM datasets.

Following the training paradigm of MAE, 75% of the image patches were randomly masked during pretraining. The model was then optimized to reconstruct the original image from the remaining 25% of visible patches, enabling it to learn high-level semantic structures from partial visual cues. We adopted the default masking ratio suggested in the MAE paper. All model parameters were initialized randomly prior to training.

Optimization was performed using the AdamW optimizer with an initial learning rate of 0.0002, scheduled via a warmup-cosine decay strategy. The learning rate was linearly warmed up from 1×10^{-7} to the base rate during the first 20 epochs. Training was conducted for 1500 epochs with a batch size of 512, using mixed precision (FP16) to accelerate convergence and reduce memory usage. The final checkpoint from the last epoch was retained as the pretrained foundation model for downstream finetuning and adaptation. Training was performed on 8 NVIDIA H800 GPUs and completed in approximately 3 days.

Standard loss and optimization objectives were employed following MAE conventions. The training loss is defined as a mean squared error between the reconstructed pixels and the original image, computed only over masked patches:

$$L_{MAE} = \frac{1}{M}\sum_{i \in M} | |\bar{x}_i - x_i| |_2^2$$

where M denotes the set of masked patches, xi is the original pixel value of patch i, and x-i represents the reconstructed pixel value predicted by the decoder. By restricting the reconstruction loss to the masked regions, the encoder is encouraged to learn semantic and structural prior rather than trivial low-level pixel statistics, a property particularly advantageous for modeling the fine-grained morphology and texture diversity inherent in EM images.

For comparative evaluation, we trained additional self-supervised baselines, DINO, MoCov3, and All4One, under identical ViT backbone configurations and hyperparameter settings. This comparison allowed us to quantitatively assess the effectiveness of the MAE-based pretraining strategy for EM imagery representation learning.

Adaptation of EMCellFound to zero-shot classification

To evaluate the generalization capability of the EMCellFound model on downstream tasks, we designed a zero-shot classification pipeline that requires no task-specific fine-tuning. In this setup, only the encoder component of EMCellFound was retained, and extract the feature embedding from encoder output. For classification, we adopted a non-parametric K-Nearest Neighbors classifier to directly perform category prediction based on the extracted embedding. Each test image x was passed through the frozen encoder to obtain its latent representation $f(x) \in \mathbb{R}^d$.

The predicted label $\hat{y}$ was then determined according to the majority label among its k nearest neighbors in the feature space:

$$\hat{y} = arg\,max \sum_{i \in N_k(x)} I\,(y_i = c)$$

where $N_k(x)$ denotes the set of k nearest neighbors of image x, y_i is the class label of neighbor i, and C is the set of all categories.

The neighborhood was defined according to cosine distance, ensuring that similarity was determined by feature orientation rather than magnitude differences. The number

of neighbors k was set as 5, which empirically balanced local consistency and global separability across categories.

No gradient update or parameter optimization was performed on the encoder during this process, ensuring a pure zero-shot evaluation of its representation quality. The classification accuracy was computed across all organelle categories to assess the encoder's intrinsic discriminative ability and semantic separability in feature space, without any supervision or task-specific adaptation

Adaption of EMCellFound to semantic segmentation

To enable effective adaptation of EMCellFound to downstream semantic segmentation tasks, we integrated a segmentation head on top of the pretrained encoder to achieve pixel-level segmentation. Given its strong performance in prior segmentation studies, we adopted UPerNet as the segmentation head to enhance both accuracy and generalization capability. The overall architecture and parameter configuration of UPerNet are illustrated in Supplementary Fig. 19. Its final layer acts as a pixel-wise classifier that outputs the predicted organelle category for each pixel. Between the pretrained ViT encoder and the UPerNet segmentation head, we inserted a lightweight neck module designed to aggregate and fuse multi-scale feature representations from both shallow and deep layers. This fusion facilitates richer contextual understanding of fine structural variations in EM images. The overall objective function combines Dice Loss and Cross-Entropy Loss, formulated as

$$L_{total} = 2L_{Dice} + L_{CE}$$

where L_{Dice} is the dice loss and L_{CE} is the cross-entropy loss.

All segmentation experiments were conducted under unified training configurations to ensure performance comparability across models.

Adaption of EMCellFound to instance segmentation

EMCellFound's support to instance segmentation tasks was achieved by appending a lightweight RTMHead to the pretrained encoder. The RTMHead module provides an efficient balance between inference speed and segmentation accuracy, making it well suited for large-scale EM datasets with densely distributed organelles. Following the standard configuration, the RTMHead was trained jointly with fine-

tuning of the EMCellFound encoder to adapt its representations towards instance-level feature discrimination. The training objective consisted of three components: a classification loss, a bounding box regression loss, and a mask segmentation loss, formulated respectively as:

$$L_{\text{ins}} = L_{\text{Focal}} + 2L_{\text{GIoU}} + 2L_{\text{Dice}}$$

where L_{Focal} denotes the Quality Focal Loss, which improves classification confidence calibration and positive-negative sample balance; L_{GIoU} is the Generalized IoU (GIoU) Loss, measuring spatial overlap quality between predicted and ground-truth bounding boxes; L_{Dice} corresponds to the Dice Loss, emphasizing accurate mask prediction for small and irregular organelle instances.

This composite loss design ensures that EMCellFound not only preserves global contextual understanding from pretraining but also acquires fine-grained instance-level discrimination required for EM image analysis.

Adaption of EMCellFound to 3D image stack segmentation

To extend the capability of EMCellFound to 3D cellular structure analysis, an efficient 2D slice-based 3D segmentation and reconstruction pipeline was developed. This pipeline leverages the strong representational capacity of the pretrained encoder to achieve high-accuracy volumetric segmentation under extremely low annotation cost. The 3D reconstruction process begins with precise cross-slice segmentation. Specifically, semantic or instance segmentation is performed on 3D image stacks by decomposing them into a sequence of consecutive 2D slices. The pretrained encoder and the corresponding segmentation heads, identical to those used in 2D tasks, are independently applied to each slice. To evaluate the few-shot adaptability of EMCellFound, the pretrained encoder and segmentation head are fine-tuned using only a very small subset of manually annotated slices. From a 3D stack containing over N (>1,000) consecutive sections, only 1-10 slices (approximately 0.1%-1% of the stack) are randomly selected and annotated. The model is then fine-tuned on these sparse annotations to generalize toward accurate segmentation on unannotated slices.

After segmentation, the 3D volume is reconstructed by re-stacking all predicted 2D segmentation masks back into their original spatial positions, yielding a voxel-wise 3D

prediction. To address potential inter-slice inconsistencies introduced by independent 2D predictions, lightweight voxel-level post-processing, including 3D morphological operations and connected component analysis, is applied to ensure structural smoothness and cross-slice continuity. The final reconstructed 3D volumes are subsequently used for quantitative organelle analysis (e.g., volume and surface area estimation) and visualization, such as volumetric rendering or surface mesh reconstruction.

User-friendly plugin interface for Napari

To facilitate the broad adoption of our foundation models by the biological community, we developed EMCFsys, a custom plugin integrated into the napari multi-dimensional image viewer (Supplementary Fig. 20 and Supplementary Video 2). The plugin was implemented in Python and encapsulates the inference pipelines of both EMCellFiner and EMCellFound within a graphical user interface. This architecture allows users to execute end-to-end workflows, including data loading, model selection, hyperparameter configuration (e.g., confidence thresholds, inference scales), and batch processing, without the need for scripting or programming expertise. The plugin supports real-time visualization of restoration and segmentation results overlaid on raw vEM data. Source code, detailed installation instructions, and tutorial documentation are publicly available on GitHub at https://github.com/yzy0102/emcfsys.

Resource availability

Lead contact

Further information and requests for resources and reagents should be directed to, and will be fulfilled by, the lead contact, Xuping Feng (fengxp@zju.edu.cn).

Materials availability

This study did not generate new materials.

Data and code availability

The EMCF dataset for the current study is available in the Science Data Bank repository at https://doi.org/10.57760/sciencedb.32271. Source data are provided with this paper. The EMCFsys plugin for napari is available on Github, at https://github.com/yzy0102/emcfsys.

References

1. Collins, L. T., Huffman, T. & Koene, R. Comparative prospects of imaging methods for whole-brain mammalian connectomics. *Cell Reports Methods* **5**, (2025).
2. Heinrich, L. *et al.* Whole-cell organelle segmentation in volume electron microscopy. *Nature* **599**, 141–146 (2021).
3. Schwarz, D. S. & Blower, M. D. The endoplasmic reticulum: structure, function and response to cellular signaling. *Cellular and molecular life sciences* **73**, 79–94 (2016).
4. Bock, A. *et al.* REEP1 Accumulation Disrupts ER Integrity and Drives Spinal Motoneuron Degeneration in Distal Hereditary Motor Neuropathy. *Advanced Science* **13**, e11483 (2026).
5. Peddie, C. J. *et al.* Volume electron microscopy. *Nature Reviews Methods Primers* **2**, 51 (2022).
6. Guo, J. *et al.* Three-dimensional analysis of membrane structures associated with tomato spotted wilt virus infection. *Plant, Cell & Environment* **46**, 650–664 (2023).
7. Wang, G., Guo, J.-S., Huang, H.-J., Zhu, Z.-R. & Zhang, C.-X. Three-dimensional interaction between Cinnamomum camphora and a sap-sucking psyllid insect (Trioza camphorae) revealed by nano-resolution volume electron microscopy. *Communications Biology* **8**, 441 (2025).
8. Li, X. *et al.* In situ architecture of the intercellular organelle reservoir between epididymal epithelial cells by volume electron microscopy. *Nature Communications* **16**, 1664 (2025).
9. Schlegel, P. *et al.* Whole-brain annotation and multi-connectome cell typing of Drosophila. *Nature* **634**, 139–152 (2024).
10. Winding, M. *et al.* The connectome of an insect brain. *Science* **379**, eadd9330 (2023).
11. Lu, C. *et al.* Diffusion-based deep learning method for augmenting ultrastructural imaging and volume electron microscopy. *Nature communications* **15**, 4677 (2024).
12. Feng, X. *et al.* Plantorganelle Hunter is an effective deep-learning-based method for plant organelle phenotyping in electron microscopy. *Nature Plants* **9**, 1760–1775 (2023).
13. Saguy, A. *et al.* DBlink: dynamic localization microscopy in super spatiotemporal resolution via deep learning. *Nature methods* **20**, 1939–1948 (2023).
14. Conrad, R. & Narayan, K. Instance segmentation of mitochondria in electron microscopy images with a generalist deep learning model trained on a diverse dataset. *Cell Systems* **14**, 58–71 (2023).
15. McCafferty, C. L. *et al.* Integrating cellular electron microscopy with multimodal data to explore biology across space and time. *Cell* **187**, 563–584 (2024).
16. Zhao, T. *et al.* A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities. *Nature methods* **22**, 166–176 (2025).
17. Ma, D., Pang, J., Gotway, M. B. & Liang, J. A fully open AI foundation model applied to chest radiography. *Nature* **643**, 488–498 (2025).
18. Conrad, R. & Narayan, K. CEM500K, a large-scale heterogeneous unlabeled cellular electron microscopy image dataset for deep learning. *Elife* **10**, e65894 (2021).

19. Heimberg, G. *et al.* A cell atlas foundation model for scalable search of similar human cells. *Nature* **638**, 1085–1094 (2025).
20. Chen, X. *et al.* Hat: Hybrid attention transformer for image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025).
21. Zhang, L., Zhang, L., Mou, X. & Zhang, D. FSIM: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing* **20**, 2378–2386 (2011).
22. He, K. *et al.* Masked autoencoders are scalable vision learners. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* 16000–16009 (2022).
23. Caron, M. *et al.* Emerging properties in self-supervised vision transformers. in *Proceedings of the IEEE/CVF international conference on computer vision* 9650–9660 (2021).
24. Chen, X., Xie, S. & He, K. An empirical study of training self-supervised vision transformers. in *Proceedings of the IEEE/CVF international conference on computer vision* 9640–9649 (2021).
25. Estepa, I. G., Sarasúa, I., Nagarajan, B. & Radeva, P. All4One: Symbiotic Neighbour Contrastive Learning via Self-Attention and Redundancy Reduction. in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* 16197–16207 (2023).
26. Cover, T. & Hart, P. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* **13**, 21–27 (1967).
27. Van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* **9**, (2008).
28. Bonifacino, J. S. & Rojas, R. Retrograde transport from endosomes to the trans-Golgi network. *Nature Reviews Molecular Cell Biology* **7**, 568–579 (2006).
29. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. in *International Conference on Medical image computing and computer-assisted intervention* 234–241 (Springer, 2015).
30. Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).
31. Cheng, B., Misra, I., Schwing, A. G., Kirillov, A. & Girdhar, R. Masked-attention mask transformer for universal image segmentation. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* 1290–1299 (2022).
32. Parlakgül, G. *et al.* Regulation of liver subcellular architecture controls metabolic homeostasis. *Nature* **603**, 736–742 (2022).
33. Xiao, T., Liu, Y., Zhou, B., Jiang, Y. & Sun, J. Unified perceptual parsing for scene understanding. in *Proceedings of the European conference on computer vision (ECCV)* 418–434 (2018).
34. Bock, D. D. *et al.* Network anatomy and in vivo physiology of visual cortical neurons. *Nature* **471**, 177–182 (2011).
35. CellMap Project Team *et al.* Near-isotropic, reconstructed volume electron microscopy (FIB-SEM) of mouse kidney (jrc_mus-kidney-2). Janelia Research Campus https://doi.org/10.25378/JANELIA.23411009 (2024).

36. Lee, W.-C. A. *et al.* Anatomy and function of an excitatory network in the visual cortex. *Nature* **532**, 370–374 (2016).
37. Lyu, C. *et al.* Rtmdet: An empirical study of designing real-time object detectors. *arXiv preprint arXiv:2212.07784* (2022).
38. Chen, Z. *et al.* Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534* (2022).
39. Zheng, Z. *et al.* Fast imaging of millimeter-scale areas with beam deflection transmission electron microscopy. *Nature communications* **15**, 6860 (2024).
40. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. in *Proceedings of the IEEE conference on computer vision and pattern recognition* 770–778 (2016).

ACKNOWLEDGMENTS

Funding

This work was supported by the National Natural Science Foundation of China (32572179). We sincerely thank Hangzhou Chengfengerlai Digital Technology Co., Ltd's support. We also thank Ben Giepmans of Universitair Medisch Centrum Groningen for the assistance and feedback.

Author information

Authors and Affiliations

College of Biosystems Engineering and Food Science, Zhejiang University, Hangzhou 310058, China

Zeyu Yu, Mengze Du, Shan Xu, Guowei Zhang, Yong He, Xuping Feng

Center of Cryo-Electron Microscopy, Zhejiang University School of Medicine, Hangzhou 310058, China

Jiansheng Guo

College of Life Sciences, Nanjing Agricultural University, Nanjing 210095, China.

Feng Liu

The Rural Development Academy &Agricultural Experiment Station, Zhejiang University, Hangzhou 310058, China

Xuping Feng

Analysis Center of Agrobiology and Environmental Sciences, Zhejiang

University.

Li Xie

Department of Computer Science, Hong Kong Baptist University, China

Bo Han

School of Agriculture, Food & Wine, The University of Adelaide, Adelaide, Australia.

Zhong-Hua Chen

Hangzhou Chengfengerlai Digital Technology Co., Ltd, Hangzhou, China.

Gaoliang Deng, Chen Rui

Author Contributions

# Zeyu Yu, Jiansheng Guo, Feng Liu, Mengze Du contributed equally to this work.

Zeyu Yu, Xuping Feng and Mengze Du conceived the study and designed the experiments. Zeyu Yu, Jiansheng Guo, Feng Liu, Mengze Du, Shan Xu, Guowei Zhang and Li Xie performed data acquisition and curation. Gaoliang Deng, Chen Rui. provided essential computational resources and storage infrastructure. Zeyu Yu and Xuping Feng wrote the manuscript. Yong He and Bo Han reviewed and edited the manuscript.

Competing interests

The authors declare that they have no competing interests.

Corresponding author

Yong He (yhe@zju.edu.cn) and Xuping Feng (fengxp@zju.edu.cn)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.docx
- supplementaryvideo1.mp4
- supplementaryvideo2.mp4