EMformer: Transformer-Based Isotropic Reconstruction for Volume Electron Microscopy

Jia He[1,2,*], Yan Zhang[3,*,†], Wenhao Sun[3,4,*], Ge Yang[1,2,†], Fei Sun[3,4,5†]

1. School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China, 100049

2. State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China, 100190

3. National Key Laboratory of Biomacromolecules, CAS Center for Excellence in Biomacromolecules, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China;

4. School of Life Scienece, University of Chinese Academy of Sciences, Beijing, China, 10049.

5. Center for Biological Imaging, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China;

* Authors contributed equally to this work

†Correspondance: Yan Zhang, yanzhang@ibp.ac.cn, Ge Yang, ge.yang@ia.ac.cn and Fei Sun, feisun@ibp.ac.cn

Abstract

Volume electron microscopy (vEM) has become the rapidly developed technique to study the 3D architecture of biological specimen such as cells, tissues, and organs at nanometer resolution, which collects series of electron micrographs of axial sequential sections and reconstructs the 3D volume, providing fruitful information of cellular ultrastructural spectrum. This technique is currently suffering from the anisotropic resolution between lateral (x, y) and axial (z) directions and the loss/damage of sections. Here, we develop a new algorithm, EMformer, based on video transformer model, to boost the axial resolution and then achieve an isotropic reconstruction of vEM. By learning the high resolution axis structures and utilizing the 3D continuity of biological structures, EMformer can recovery the axial information and repair the random loss/damage sections based on self-supervision strategy, achieving a higher resolution than existing methods, which is validated for both simulated FIB-SEM dataset and experimental ssTEM dataset. Besides visual validation, the segmentation efficiency and statistical precision of various ultrastructures, e.g. neurons, mitochondria, vesicle, and membrane bilayers, also prove a better performance of EMformer. Therefore, using EMformer, we could achieve isotropic reconstruction from anisotropic axial sampling, which will increase the throughput of vEM to study a large scale of biological architectures.

1. Introduction

In recent years, with the continuous development and optimization of electron microscope imaging technology, big data storage, image processing technology and computer hardware technology, volume electron microscopy(vEM) has developed rapidly. vEM allows directly depict

3D ultrastructure of organisms, cells[1-4] and tissues[5,6], especially has been widely used in connectomics studies of neurons [7-10]. The physical scale studied by vEM is also getting larger and larger, from the nanoscale to tens of microns and even hundreds of microns [9,11]. Thus, fast development of vEM has promoted its widely application in life sciences [12,13], medicine [14-16] and clinical diagnosis [17-20], and become one of the most important and state of the art technologies in the above fields. For these reasons, vEM was named one of Nature's top 7 technologies to watch in 2023.

Even with the most advanced technology, there are still difficulties or flaws that need to be overcome for vEM. According to different imaging methods, vEM technology is divided into volume transmission electron microscopy (TEM) and volume scanning electron microscopy (SEM)[9]. The volume TEM technology includes serial section electron tomography (ssET) and serial section transmission electron microscopy (ssTEM). Volume SEM techniques include serial block surface scanning electron microscopy (SBF-SEM)[11], serial ultra-thin section scanning electron microscopy (ssSEM) and focused ion beam scanning electron microscopy (FIB-SEM)[11]. Actually, both of ATUM-SEM [21,22] and AutoCUTs-SEM [23] which can automatically tape-collect serial sections are ssSEM technology.

For vEM, the Z-axis direction of the reconstructed volumes of ssTEM and ssSEM [11] can only contain the integrated information of the thickness of each session, so the session thickness determines the axial voxel size. Usually, section thickness is around tens of nanometers, which is several times or even tens of times the lateral plane, this results in the anisotropy of the information between the lateral and axial directions. Similarly, the axial resolution of SBF-SEM depends on the thickness of the sample sequentially cut and the depth that electron beam penetrates. FEI designed Volume Scope with an automatic deconvolution and multi-energy imaging method to further enhance the axial resolution of SBF-SEM. But at present, the isotropy resolution achieved by Volume Scope is at the level of dozens of nanometers. Therefore, ssTEM, ssSEM and SBF-SEM are all anisotropic vEM techniques if the goal is ten nanometers or even higher. ssET and FIB-SEM can achieve isotropic Volume EM reconstruction. However, because each session of ssET can only be tilted at a limited angle range, its reconstructed result is not complete in Fourier space with information lost in high angle range, which is known as the missing cone problem. At present, none of the methods can fill missing cone with complete accurate high-frequency information. On the other hand, ssET can only be applied to those small volumes ranging from hundreds of nanometers to several microns. FIB-SEM can control the cutting thickness more precisely by controlling ion beam current and accelerating voltage, so as to realize isotropic 3D reconstruction. C. Shan Xu and Harald Hess have developed enhanced FIB-SEM that enables large-scale isotropic reconstruction, with cumulative collective volumes up to 10^6 microns in several months [24]. However, the time cost is a big issue which must be considered for large volume 3D reconstruction. And on the other hand, enhanced FIB-SEM is only available in few institutions. Therefore, FIB-SEM is currently mainly applicable to the isotropic reconstruction of small volume samples. Further development of these above techniques is needed for efficient reconstruction of isotropic electron microscopy with large scale and high resolution.

Our research goal is to routinely achieve isotropic resolution of 20nm or higher for large volume samples with the above vEM technology. Due to the limitations of the above techniques, large-scale and high-resolution isotropic reconstruction cannot be widely realized. Previously, people usually made interpolation to the reconstruction result to reduce the anisotropy effect

caused by the loss of axial information. But the error and artifact caused by interpolation are obvious, especially in the case of anisotropic reconstruction with high anisotropic factor such as 8 ×, 10×, even larger.

In recent years, various types of deep learning algorithms have continuously emerged and promoted the development of artificial intelligence. Subsequently, its application in vEM has gradually increased. Heinrich first introduced the concept of deep learning to vEM in an attempt to solve the anisotropic reconstruction problem [25] , showing the feasibility of 3D-FSRCNN and 3D-SRUNet to perform isotropic reconstruction by supervised learning and the result is superior to bicubic interpolation. But that was a supervised 3D reconstruction method which only carried out in the case of having ground truth data. However, the isotropic ground truth volumes of ssSEM/ssTEM/SBF-SEM can't be obtained in vEM studies, which limits the application of such supervised deep learning methods in vEM. Self-supervised deep learning methods without the need of ground truth are more suitable for isotropic reconstruction of vEM, including Kausumi Hagita et. al.[26], Xue Yang Fu et. al.[27], Hyoungjun Park et. al.[28] et. al. These methods are all based on generating adversarial networks (GAN), using high-resolution two-dimensional images in the lateral direction as unpaired self-supervised signals, and guiding the model to restore axial low-resolution images by domain transferring. Specifically, Kausumi Hagita et. al.[26] firstly used SRGAN[29] to restore simulated anisotropic FIB-SEM volumes and Hyoungjun Park et. al.[28] applied similar methods to fluorescence 3D imaging. Xue Yang Fu et. al.[27] adopted a better CycleGAN[30] to cope with the blind degradation of vEM and performed isotropic reconstruction on real ssTEM data with 10× anisotropy factor. However, these GANs still adopt a 2D network architecture, which has shortcomings in modeling the inter-session continuity of 3D volumes. However, due to the high difficulty and instability of training, the final reconstruction results may contain many artifacts, such as texture artifacts or structural deformations. Some work also uses video models to recover axial information, such as Zejin Wang et. al. [31] designed a video frame interpolation model with self-attention mechanism, completing intermediate session between two sessions, but only achieve 2×axial information recovery.

More recently, diffusion model-based vEM isotropic reconstruction methods have emerged, such as DiffuseIR[32], EMDiffuse[33]and DiffusionEM[34]. Overall, the diffusion model has better high-frequency information recovery ability and training stability compared to GAN. Among these methods, DiffuseIR and DiffusionEM adpot zeroshot restoration paradigm, training the diffusion model using lateral sessions, and then restore the anisotropic axial sessions during sampling. EMDiffuse uses diffusion model to implement a toolbox for electron microscopy restoration, including super-resolution, denoising, and isotropic reconstruction. However, these diffusion model methods still use 2D model architecture, only partially compensating for the lack of spatial continuity through positional encoding. Moreover, the training and sampling of diffusion models are time-consuming, which becomes a bottleneck problem in practical applications.

Another factor affecting the effectiveness of vEM reconstruction is artifacts such as section damage, creasing, contamination and other human defects which are inevitable during sample preparation. This not only affects the pre-processing operations such as the stitching and alignment between the images, but also reduces the overall spatial resolution of the subsequent reconstruction results. For samples with a lot more defects, samples can only be re-prepared to obtain high-quality images, which slows down the speed and efficiency of vEM research. In recent years, image inpainting method [35-39] by deep learning has been developed rapidly in the fields of

image editing and photo restoration, making up for the disadvantages that the information repaired by traditional methods does not match the original semantics. Moreover, these technologies are also gradually developed and applied in medical image repair such as MRI spectroscopy[40]. Commonly used inpainting networks include convolutional neural networks based on context encoders[41], full convolutional neural networks based on global and local context discriminators[42], vanill GAN networks based on generating images with random noise[43], U-Net networks based on pyramid context encoders[44]，transformer network[45] and some other related methods[46-49]. To our knowledge, the application of inpainting in vEM is still very rare. By using the continuity priori of the ultrastructure between continuous session or continuous block face images, and combining with the long-term attention mechanism of transformer model, we believe that the transform network can be well extended to the inpainting of vEM images with defects such as missing sessions, damages, wrinkles, et. al. This will not only repair those above defects in vEM raw images, but also save the time and economic cost for sample preparation.

In this work, we propose the EMformer deep learning model, which uses a video transformer to solve the isotropic reconstruction problem of vEM, and also handles image inpainting for defect of xy planes. EMformer restores axial low-resolution information by learning the high-resolution lateral information of anisotropic volumes in a self-supervised manner. By adopting a video transformer architecture that includes self-attention and mutual attention mechanism, the spatial continuity of 3D structures can be effectively learned, avoiding the discontinuous layered artifacts in 3D reconstruction volumes that typically appear in 2D models. Moreover, for data with session deficiency even session lost during data collection, EMformer can be extended to EMformer+, performing session image inpainting and isotropic reconstruction simultaneously. For example, EMformer+ can achieve reliable session inpainting based on inter-frame continuity and video transformer model. In addition, EMformercan achieve arbitrary-scale isotropic reconstruction during inference phase, which enables EMformer easily transfers on other vEM data with different anisotropic factors, even for fractional factors. Pretrained EMformer has good transferring robustness, indicating its potential as a universal pre-trained isotropic reconstruction model for vEM. Both on simulated anisotropic FIB-SEM and real anisotropic ssTEM data, EMformer has significantly improved the morphology restoration of organelle ultrastructure with different types and scales, such as neuron, mitochondria, vesicles and bilayers. Based on multiple qualitative indicators including reconstruction accuracy, artifact restriction, training stability and speed, EMFormer is far superior to interpolation, better than existing self-supervised methods, and even close to existing supervised methods. It further facilitates downstream analysis related to connectiomics or structures, including 3D segmentation and statistical analysis.

2. Results

For isotropic reconstruction problem of vEM, we designed a video transformer based self-supervised deep learning model named EMformer (Section 2.1), which shows superior isotropic performance than traditional interpolation algorithms and other self-supervised algorithms both for simulated anisotropic FIB-SEM data (Section 2.2) and real anisotropic ssTEM data (Section 2.3). The FSC criterion is used to give the accurate resolution of reconstructed volume by EMformer. Experiment on simulated anisotropic data demonstrate that for 40nm session thickness and 5nm lateral spatial sampling, the axis information above 18nm can be accurately recovered. The

segmentation of organelle such as neurons, mitochondria and corresponding statistical analysis further demonstrate the accuracy and reliability of EMformer. Moreover, EMformer can achieve isotropic reconstruction with any anisotropic factor and has good transferring robustness in various vEM data, such as different tissues and organs collected by FIB-SEM, ssTEM or ssSEM (Section 2.4). EMformer can be extended as EMformer+ when facing common session deficiency, such as performing session inpainting for missing sessions, contaminated or broken areas while isotropic reconstruction (Section 2.5). According to our current knowledge, EMformer is the first method to realize isotropic reconstruction and inpainting simultaneously.

2.1 Algorithm design and network architecture

Since the isotropic ground truth of vEM cannot be obtained, we adopt a self-supervised learning strategy for anisotropic volume. For dataset preparation, we crop the anisotropic volume into sub-volumes with specific size to generate training/validation/test dataset (Fig. 1a). The training set usually accounts for 60%~80%, the validation set is the remaining area that does not overlap with the training set, and the test set can be any other data that is different from the training set. In order to utilize lateral high-resolution information for self-supervision, we perform anisotropic degradation to sub-volumes in training set along X/Y axis to construct paired training data. The degradation process is performed by anisotropic average pooling operator(Fig. 1a), which is suitable for multiple modalities of vEM (Supplementary Note 1) and used in previous work[25]. During training, the paired sub-volumes are augmented with eight types of orthogonal rotation, and regarded as video clips along Z axis (Fig. 1b). After feeding them into EMformer, the sub-volumes are super-resolved along simulated degraded axis (X/Y), and the weighted sum of L1 and SSIM difference is taken as loss function (Fig. 1b). During inference, the input sub-volumes are also augmented with eight types of orthogonal rotation, and then regarded as video clips along X/Y axis (Fig. 1b). After model inference, 8 super-resolved results of rotation augmentation are average ensembled to generate final reconstruction. Finally, the reconstructed sub-volumes in test dataset are stitched according to the cropping coordinates to obtain the whole isotropic reconstructed volume.The network architecture of EMformer (Fig. 1c) is designed as multi-scale symmetrical stack of transformer blocks with an arbitrary-scale up-sampling module. Specifically, given a sub-volume or video clip, the process flow is as follows: Firstly, feature extraction is performed by spatial 2D convolution for each frame in video clip. Then, the feature maps are processed by multiple transformer stages, where 1~7th stages perform multi-scale symmetrical feature reconstruction in Unet[50]shape, and 8th stage perform fine feature fusion at the original scale. The up-sampling and down-sampling are symmetrical in layout and operated before the feature maps entering a certain transformer stage, only changing the spatial size of each video frame while keep temporal length unchanged. Each transformer stage contains stack of transformer blocks to jointly extract features, perform the motion alignment between frames, and fuse the inter frame information. The number of transformer blocks in 8th stage is usually more than 1~7th stages (M>N). Finally, the feature maps are isotropic reconstructed along simulated degraded axis (X/Y) and to generate isotropic sub-volume. Global residual connection and skip connection are adopted to better reuse features and facilitate network learning.

Inspired by video restoration transformer (VRT) [51], the transformer block includes two types of attention modules, MSA (Multi-head Self Attention) and MMA (Multi-head Mutual Attention),

to perform feature extraction and motion alignment (Supplementary Note 2). Self-attention[52] (Extended Data Fig. 1a) uses the self-similarity prior of images to make the feature focus more on important regions, which has been proved to be effective in various computer vision tasks[53] As shown in Extended Data Fig. 2, the self-attention map shows which areas of input video clip are focused on, illustrating the long-distance dependency modeling ability across frames and the robustness to different structures. Compared to self-attention, mutual attention (Extended Data Fig. 1b) emphasizes joint motion estimation and implicit feature alignment between two adjacent frames. The mutual attention map reflects the position correspondence between these two frames. Compared to explicit motion estimation and compensation method such as optical flow[54,55]and image warping, mutual attention has simpler implementation, more robust to large structural deformation between frames avoiding mismatch artifacts, and more flexible for information warping in latent space. Multi-head strategy[52]is used for the above two attention mechanisms to enhance semantic representation capabilities. To reduce computation cost of video processing, it is necessary to divide attention windows and shift them (Supplementary Note 2) in adjacent network layer to preserve long-distance dependency [56], as shown in Extended Data Fig. 1c.

After multi-scale transformer stages, the feature maps are isotropic super-resolved by reconstruction module, composed by anisotropic pixel shuffle[57]and trilinear interpolation to realize arbitrary scale up-sampling. Firstly, the pixel shuffle is used in anisotropic way, only enhancing the resolution along single spatial axis direction. Secondly, the followed trilinear interpolation enables plug-and-play arbitrary scale up-sampling, even for fractional scale factor. It should be noted that anisotropic pixel shuffle can up-sample at any positive integer scale factor, as long as the number of channels increases linearly with the scale factor. But it restricts that the scale factor during model inference has to consistent with the training phase. By following trilinear interpolation after anisotropic pixel shuffle, it enables model infer on a scale factor that is inconsistent with training phase, facilitating the pretrained model transfers to other vEM data with different anisotropic scale factors, even for fractional number. Besides, trilinear interpolation is effective for back propagation and lightweight without increasing parameters or computation.

In summary, EMformer integrates feature extraction, motion alignment, and information fusion at multiple resolutions by multi-scale video transformer stages and perform arbitrary scale up-sampling by combination of anisotropic pixel shuffle and trilinear interpolation. Comparing with existing isotropic reconstruction methods for vEM, EMformer can trace structure movement between frames and maintain long-distance continuity, while 2D deep learning models can't perform well. It also avoids training instability and high-frequency artifacts that often occur in GAN-based model such as SRGAN[29] and CycleGAN[30]. It achieves excellent performance far beyond the bicubic interpolation and close to supervised method [25] (Fig 2).

Fig. 1: Algorithm design and process framework of Emformer.

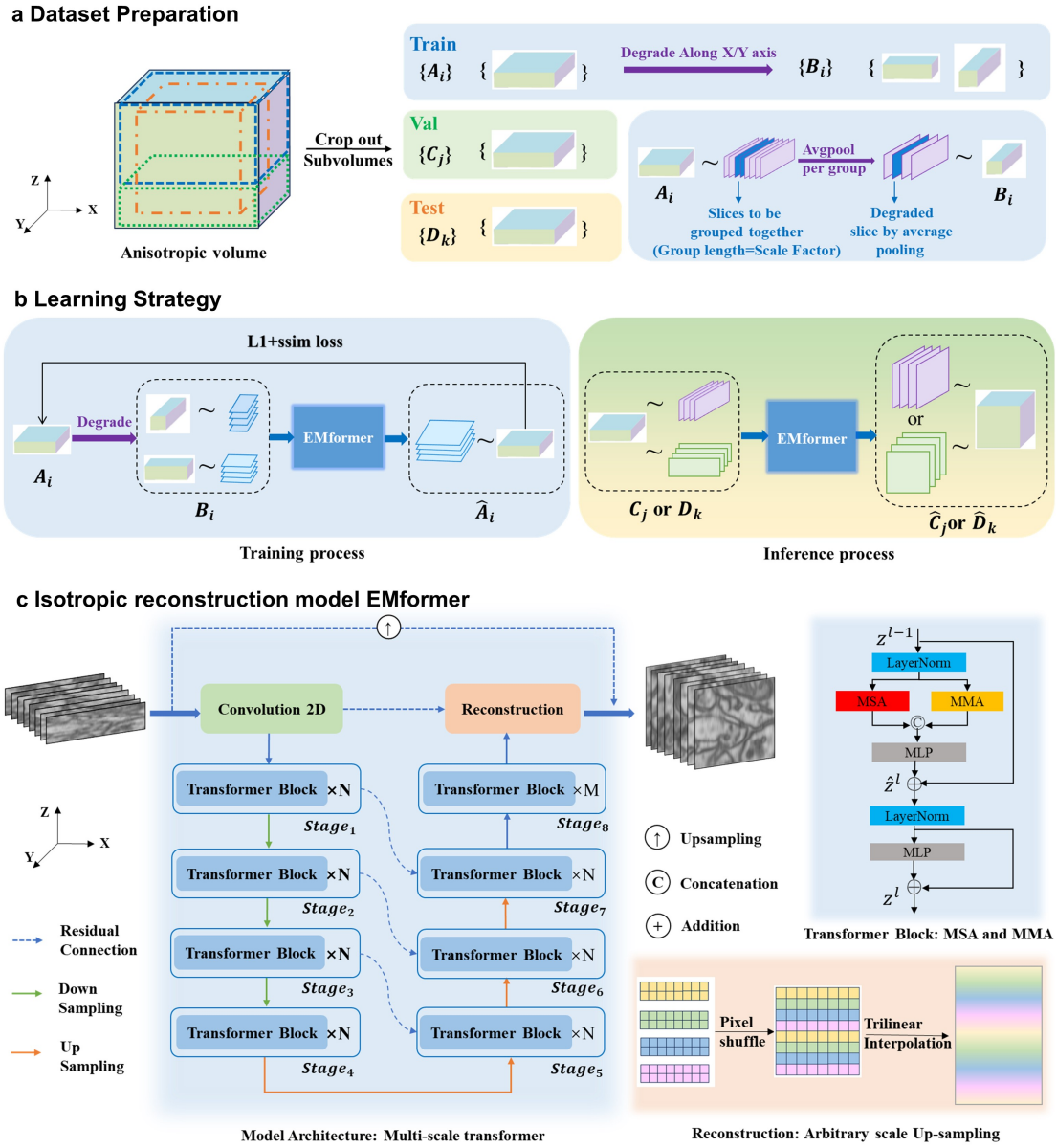


Fig. 1 | Algorithm design and process framework of Emformer. a, Generating sub-volume sets for deep learning from raw anisotropic vEM volume. Using anisotropic average pooling as degradation simulation method, the sub-volumes are degraded along the X or Y axis to generate paired LR-HR data for self-supervised training. A_i, C_j, D_k represents the training set, validation set and test set respectively; B_i represents the degraded volume along X or Y axis. **b,** In the training and inference process, the model regards input sub-volume as video clip along Z axis and X/Y axis respectively. Loss function is set as the weighted sum of L1 and SSIM loss. C_j *or* D_k means the input data which needs to be inferred by EMformer network. $\hat{A}_i$ is the output of the EMformer network, which is super-resolved along degraded axis. $\hat{C}_j$ or $\hat{D}_k$ is the inferred volume. **c,** Model architecture is designed as a multi-scale video transformer and up-sampling reconstruction module. The 1st to 7th stage are made for multi-scale feature extraction, and each scale have N transformer blocks containing MSA and MMA to perform implicit motion alignment. The 8th stage has M transformer blocks designed for feature fusion. Arbitrary anisotropic restortaion is achieved by trilinear interpolation after pixel shuffle inside the network.

2.2 Isotropic reconstruction for simulation dataset with ground truth

We downloaded isotropic data of mouse brain neurons collected by FIB-SEM from EPFL website (https://www.epfl.ch/labs/cvlab/data/data-em/) and simulated 8× anisotropy by averaging multiple frames every 8 sessions. The sample contains not only subcellular structure such as neurons and mitochondria, but also the ultrastructure of vesicles and cell boundaries at the nanometer level. It is a typical sample for ultrastructure research using vEM technology, so we use this data to simulate serial session generation and test the effectiveness of EMformer algorithm. Comparison with the supervised method[25] and CycleGAN[27], we found that EMformer presents the best isotropic information in both image space and Fourier space (Fig. 2a). The power spectrums of x-z planes show that EMformer can restore Fourier space information within the same diameter as supervised network; while both CycleGAN[27] and bicubic interpolation present elliptical diffraction information in the Fourier space (Fig. 2a). More intuitively, the reconstruction results of the x-z planes of EMformer method showed the same details as the supervised method even the same with ground truth (first three rows of Fig. 2a) such as the bilateral membranes (red box) and the ridge mitochondria (blue box). However, although the reconstruction of CycleGAN and bicubic interpolation also has some restoration of axial information, the overall contrast is blur and the structure details such as bilateral membrane cannot be presented.

In order to further confirm the effectiveness of EMformer, we further segmented and extracted mouse neurons and mitochondria and performed morphological comparative statistics (Fig. 2b, c and d). For the sake of generality, we segmented three different locations of neurons in the reconstruction results of each method. The neurons recovered by EMformer had no jagged edge, and the voxel coincidence were closest to the ground truth (Fig. 2b and d). Except supervised methods, the segmentation breakpoints of neurons in the EMformer reconstruction results are also the fewest among all deep learning-based methods and interpolation method (Fig. 2c and d). For neuron 2 and 3，the number of broken points is even less than supervised method. We believe that this is because EMformer makes full use of the 3D spatial information of the data through the video transformer model, so the reconstructed results are more continuous both in the whole volume and local ultrastructure such as synapse. The mitochondrial segmentation for EMformer reconstruction shows easily recognizable and complete mitochondria. Compared with CycleGAN or interpolation methods, the mitochondria in the reconstructed results of EMformer have the highest colocalization volume with isotropic ground truth by FIB (33.76μm^3), almost the same as that of supervision method (33.87μm^3), while the interpolation method has the smallest colocalization volume (33.19μm^3) (Extended Data Fig. 3).

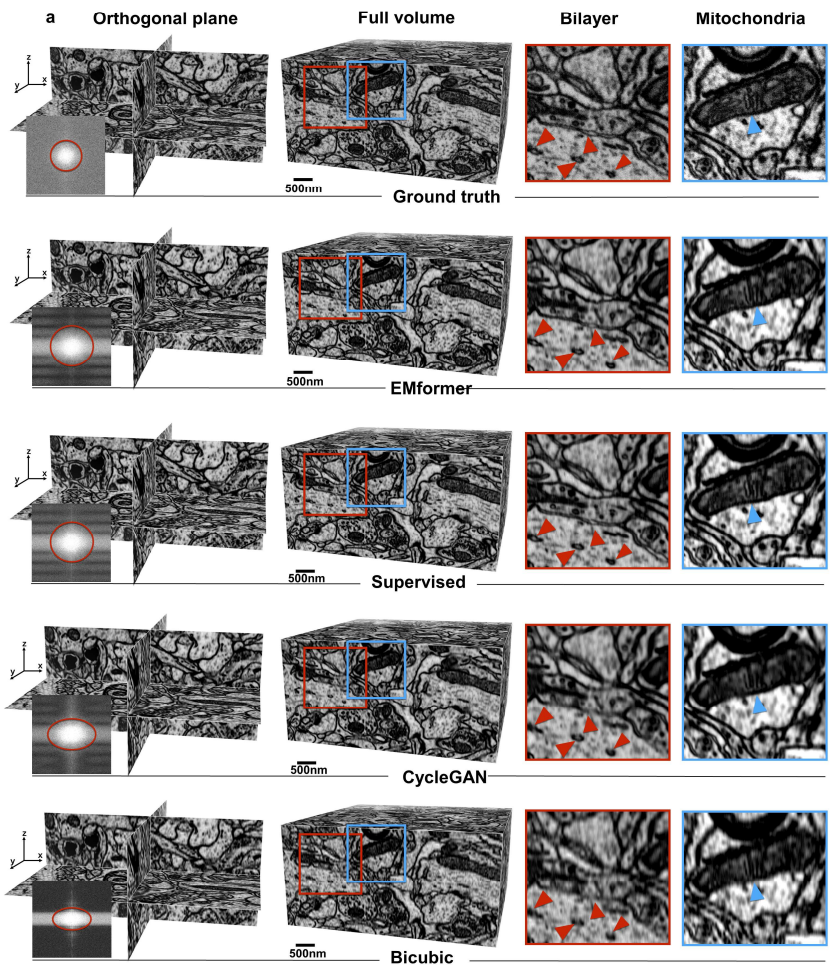
FSC (Fourier shell Correlation) has been used widely as a golden standard for reliable reconstruction resolution estimation in cryo-EM. Here, we introduced FSC calculation into this EPFL data with ground truth. FSC of isotropic reconstruction result was calculated by comparing the Fourier shell cross-correlation with ground truth volume. Except for supervised deep learning method, FSC performance of EMformer is significantly better than bicubic interpolation and CycleGAN methods (Fig. 2e). By FSC-0.5 criterion, EMformer could reach 18nm isotropic resolution. This open EPFL data has 5nm isotropic resolution, and we simulated the axis degradation with anisotropic 8 which means the session thickness simulated as 40nm. The 18nm recovered information imply us that if people want to get 20nm axis resolution, slicing sample with 40nm thickness is completely replace the 20nm slicing stratagem. This will save half the

number of sessions, and further saves at least 50% of the time in the entire process such as sample preparation, 2D registration and 3D reconstruction. Of course, at the same time, the electron microscopy and computer expenditure costs are saved.

We also calculated the performance of EMformer with different anisotropic factor and quantify the restoration quality by pixel-based quantitative indicators PSNR, structure similarity SSIM and MSSSIM, visual perception indicator LPIPS for both whole 3D reconstruction volume and three orthogonal planes (XY, XZ, YZ) respectively. Either anisotropic factor 4 or 8, all the above metrics for EMformer are better than the methods based on bicubic interpolation or CycleGAN network (Extended Data Fig. 5). This is consistent with the fact that the ultrastructure including neuron (Fig. 2) and mitochondrial (Extended Data Fig. 3) presented by EMformer is closest to the ground truth.The double-layer membrane structure on cell boundary region recovered by EMfomer is even more continuous than that recovered by supervised method (Fig. 2a, red arrows).

All the above analysis demonstrates that EMformer can recover ultrastructure with higher-resolution information in organelles and achieve reliable isotropic reconstruction results with higher FSC, lower reconstruction error and better visual perception (Fig. 2, Extended Data Fig. 3 and 4).

Fig. 2: Comparison of isotropic reconstruction results of EMformer with existing methods on EPFL dataset with ground truth.



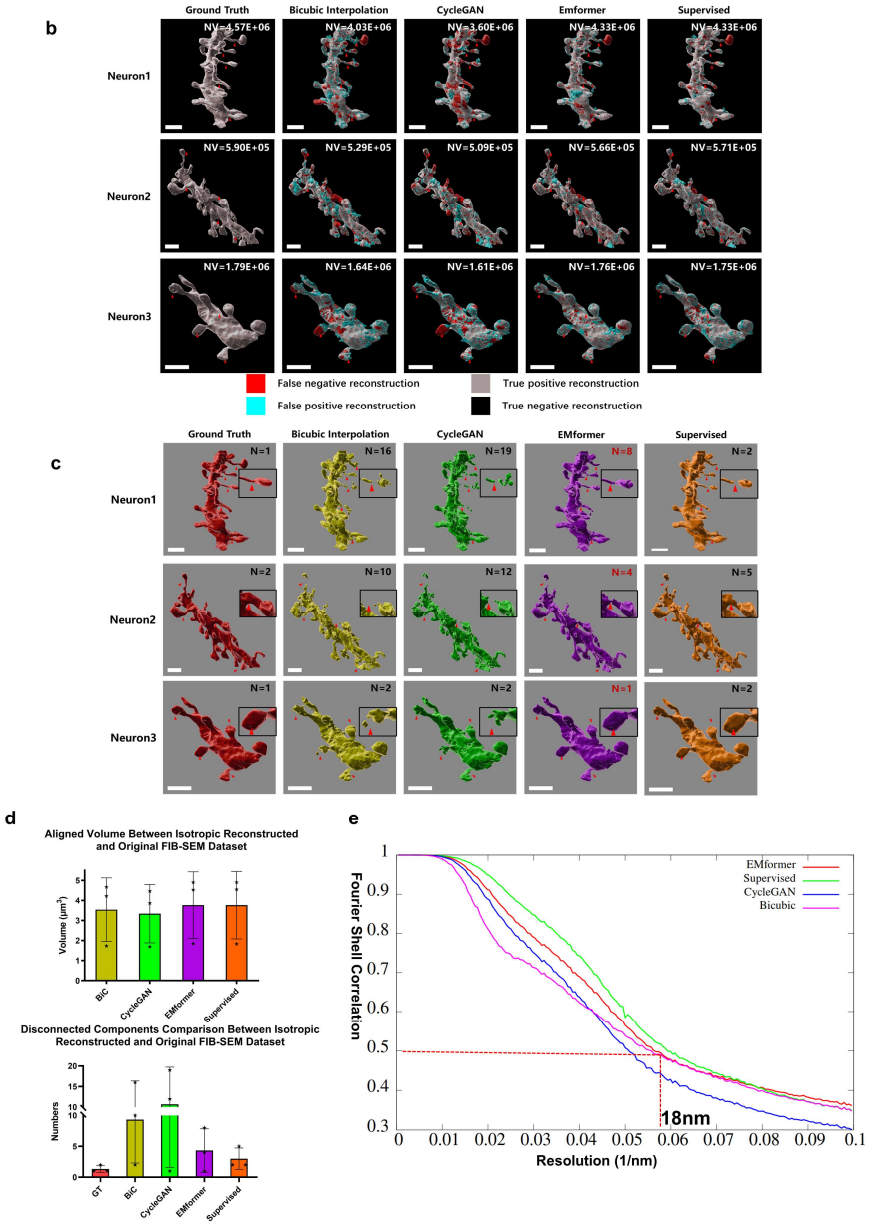


Fig. 2 | Comparison of isotropic reconstruction results of EMformer with existing methods on EPFL dataset with ground truth. a, Each line from top to bottom represents ground truth, the reconstruction of EMformer, CycleGAN and bicubic interpolation respectively. Each method was shown with orthogonal view and full volume, while the power spectrum of the orthogonal x-z plane are presents, and the local membrane structures and mitochondria from the full volume are highlighted with enlarged image and arrows (red arrows for membrane structures and blue arrows for mitochondria) **b,** Three neurons from the results of different reconstruction methods were segmented. The colocalization analysis between reconstructed neurons by different methods and ground truth neurons were carried out. The values in the figure represent the volume of the colocalization in cubic microns. Except for supervised method, the neurons segmented from EMformer reconstruction result have the largest colocalization volume with the ground truth. **c,** Analysis of neuronal connectivity. The values in the figure represent connectivity of neurons. A larger value indicates more breakpoints and poorer connectivity. Among the neurons isolated from 4 different isotropic reconstruction results, the interpolation method has the worst neuronal connectivity. Neuron connectivity of our EMformer method is better than CycleGAN. It's even also better than supervised method for neuron 2 and 3. **d,** The volume of overlapping regions and the number of discontinuous

breaks of neurons in Fig. 2b and c were statistically analyzed. As shown in Fig. 2b and Fig. 2c, bicubic interpolation method has the most breakpoints, EMformer method is superior to CycleGAN method, and breakpoints for neurons from ground truth are the least. The colocalization overlapping volume of neurons for the interpolation method is the smallest, while the colocalization volume for EMformer reaches the same level as supervised methods. **e,** FSC cross correlation calculation between volume of each isotropic reconstruction method and the ground truth volume. With FSC 0.5 standard criterion and under the condition of 5 nm imaging pixel in x-y plane, EMformer can recover reliable axial information at 18nm. Moreover, FSC curve for EMformer (red curve) had significantly higher correlations with ground truth structure, second only to the FSC curve of the supervised approach （green curve）.

2.3 Isotropic information recovery by EMformer for real data without ground truth

To further demonstrate the effectiveness of EMformer and its applicability to other SEM modal data, we further tested the public ssTEM dataset Cremifrom website(http://cremi.org/). Cremi provides publicly available raw ssTEM images of the Drosophila melanogastrii brain for algorithmic development and evaluation of neural circuits.This Cremi data is collected by ssTEM with anisotropic factor 10 and there is no ground truth to make comparison. Therefore, we directly compare with traditional bicubic interpolation methods and CycleGAN networks.The results show that for such high anisotropy data, EMformer reconstruction has the best structural details (Fig. 3a and b). The results of CycleGAN network can recover some axial information, but it can't recover the details of membrane like structures such double-layers of membrane (Fig. 3b). The result of the bicubic interpolation method is undoubtedly the worst, the overall axial structure is fuzzy, and the structural details are smoothed out (Fig. 3b). On the other hand, the whole volume generated by EMformer demonstrate continuous and nature structure information in 3D due to space continuity learning advantage of video transformer in EMformer; while the volume obtained by CycleGAN or bicubic interpolation has obvious layer artifacts in z axial direction. The reason is that CycleGAN is a two-dimensional model, frame-by-frame inference along x(y) axis generates layered discontinuous artifacts on yz(xz) planes, which are difficult to eliminate even in the presence of training data augmentation. On the contrary, EMformer kept each session as a continuous video frame and adopted the self and mutual attention mechanism for inter-frame motion alignment, thus maintained the spatial continuity of the reconstructed volume.

Fig. 3: Comparison on real ssTEM data Cremi with anisotropic factor 10.

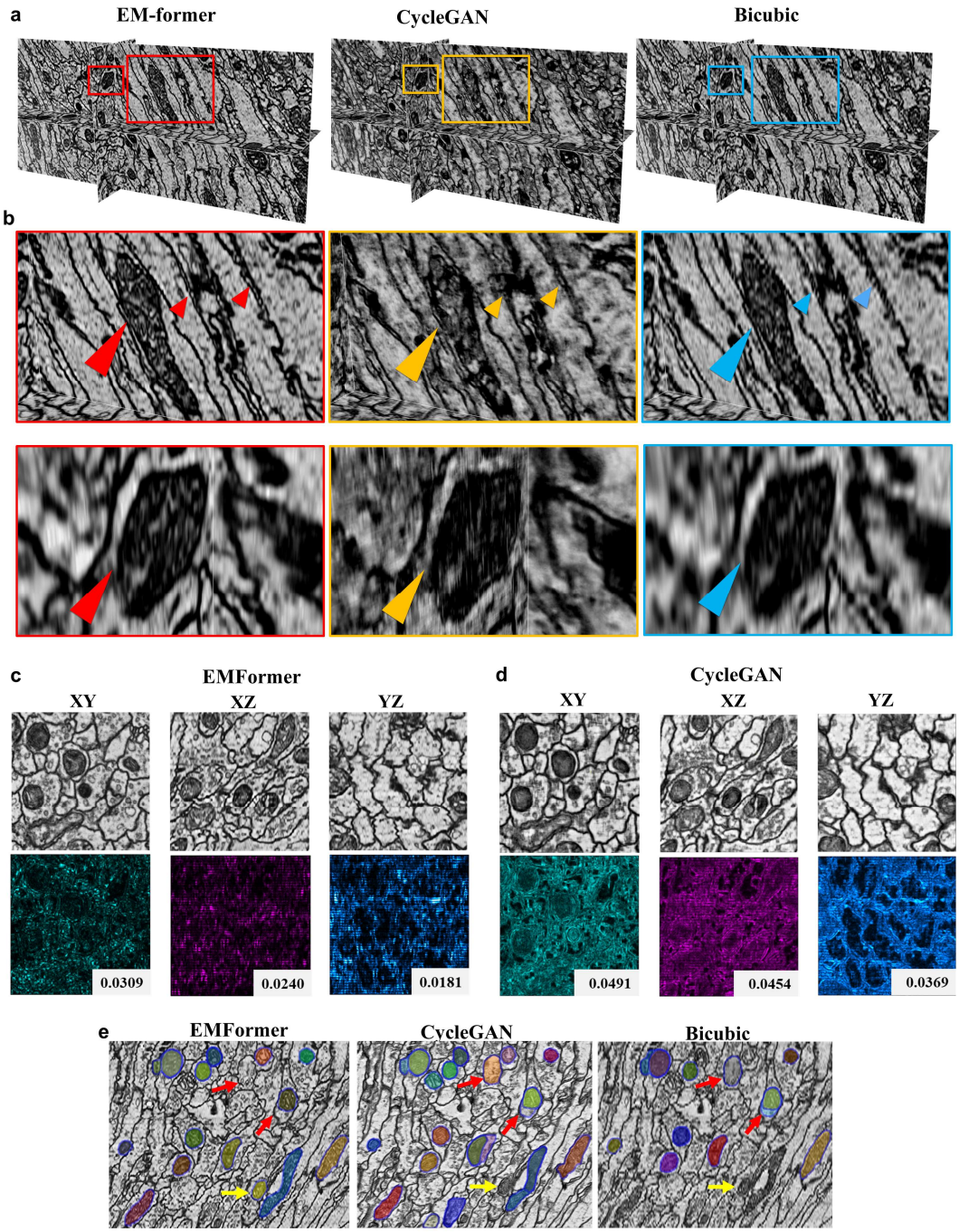


Fig. 3: Comparison on real ssTEM data Cremi with anisotropic factor 10. a, EMformer is clearer and sample details are more natural, while bicubic interpolation produce blurred structures and CycleGAN exhibit layered artifact. **b,** Highlighted regions for above three reconstruction results in A. EMformer (red rectangle) shows more complete and clear structure at the boundaries of the membranous structure and the structure of the internal ridges of the mitochondria compared to CycleGAN (yellow rectangle) and bicubic interpolation (blue rectangle). **c,d,** Uncertainty map comparison for EMformer and CycleGAN. The artifact signal of CycleGAN is significantly larger than EMformer in three orthogonal directions, which means the predictions of 8 kinds of rotation augmentation has larger standard deviation. EMformer can reconstruct stable and reliable structural information than CycleGAN. **e,** Mitochondria Segmentation using SAM for reconstruction result of EMformer, CycleGAN and bicubic respectively. Mitochondrial segmentation is performed using the instance segmentation method in SAM.

To further evaluate the reliability of model prediction on Cremi dataset without ground truth, artifact analysis was performed based on model uncertainty [33]. Training data was augmented based on 8 orthogonal rotations during model inference, so model uncertainty can be measured by standard deviation of those 8 predictions, and uncertainty map of this deviation was generated to represent reliability of each voxel position during model prediction. This is a disguised statistical verification method for the reality constraint when there is no truth value to compare. We compared the uncertainty map of EMFormer and CycleGAN on Cremi dataset along each orthogonal axis (Fig. 3c), it shows that CycleGAN has significantly more artifacts than EMformer, although they have similar structure texture. To verify the correlation between uncertainty map and error map, we evaluate on EPFL dataset with ground truth (Extended Data Fig. 4). Both the uncertainty maps and error maps have the same distribution for bright region that mainly appear at membrane boundaries and complex structures. It should be noted that for the defective area, it has been removed by network, so it is not reflected in the uncertainty map, but still exists in the error map.

We tried to segment the reconstruction result with "segment anything model" (SAM)[58]. Compared with the results of CycleGAN (Extended Data Fig. 6c and d) or bicubic interpolation (Extended Data Fig. 6d and e), the segmentation results for EMformer are obviously better (Extended Data Fig. 6a and b). The boundaries of vesicles (Extended Data Fig. 6a) and membrane structures (Extended Data Fig. 6b) were clearly separated. As for CycleGAN, due to the presence of artifacts, the edge segmentation of vesicles is not accurate, and the double layer membrane is predicted to be a single layer membrane. The mitochondria segmentation of EMformer was also more accurate, and segmentation results for CycleGAN or bicubic method show that some mitochondrial was not recognized (yellow arrow), while some area which is not mitochondrial was misidentified (red arrow) (Fig. 3e). This is because EMformer could reconstruction continuously for 3D space information, especially the axial information was restored, the staircase artifacts were removed. Therefore, the reconstruction of EMformer facilitates downstream analysis of ultrastructure and the study of biological mechanisms., as well as over segmentation of mitochondria.

2.4 The transferring robustness of EMformer

For the isotropic reconstruction of vEM, there are two problems between the scale factor and the model. One is that different data have different anisotropy scale factor, which may be positive integers or decimals. The other is how to ensure that the model trained on the data with anisotropy a can be transfered to another data with anisotropy $b(b \neq a)$. We solve the above two problems through arbitrary scale upsampling module, which contains anisotropic pixel shuffle followed by trilinear interpolation (see Section 2.1 for details). This will allow EMformer to adapt to data with different anisotropic factors and make pretrained model transfer directly among them. We performed model inference on arbitrary scale, using the model trained on 10x anisotropic Cremi dataset used in Section 2.3. For CycleGAN, to realize arbitrary scale reconstruction during inference without changing model weights, we simply change the scale factor of up and down sampling module in the CycleGAN network. Compared with the CycleGAN network and bicubic interpolation, EMformer obtained complete and clearer ultrastructure information, CycleGAN shows obvious artifact due to the mismatch between network weight and anisotropic factor;

Structures by bicubic interpolation become blurred for large anisotropic factors (Fig. 4a). We also evaluate on EPFL dataset with ground truth (Extended Data Fig. 7) by simulating anisotropic data of any scale factor with average pooling operation, as described in section2.1. Then the 5nm isotropic data is reconstructed by different methods. The PSNR, SSIM, et. al. validation indicators of these method are shown as Supplementary Table 2. EMformer obtain the best evaluation indicators among all the three methods and different anisotropic factors.

The plug-and-play arbitrary scale up-sampling enables EMformer to transfer between different anisotropic data, including collected by different electron microscopies, without the need for training on different data each time. For examples, data collected with FIB-SEM was simulated as 8x anisotropic degradation along z axis, it can be recovered from the model trained on Cremi data collected by ssTEM with anisotropic factor 10, and vice versa (Fig. 4b). The performance of model fine-tuning across datasets is comparable to that of direct training and inference in the target domain data, which shows that the model transferring has good robustness. In this way, once the pre-trained model is available, it can spend less time to achieve good reconstruction through transfer learning in practical applications. (Fig. 4 b).

Fig. 4: Model inference with different anisotropic factors and transferring robustness.

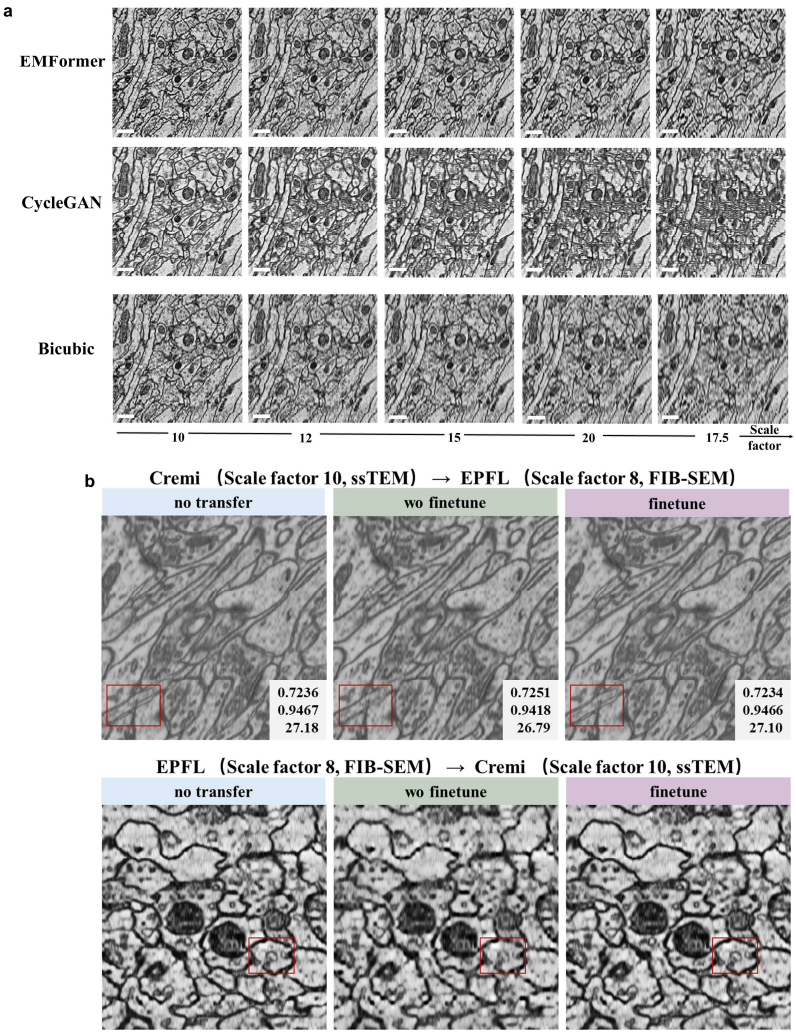


Fig. 4 | Model inference with different anisotropic factors and transferring robustness. a, We directly apply EMformer trained on 10x anisotropic Cremi dataset to simulated anisotropic Cremi data with arbitrary factors of 10x, 12x, 15x, 17.5x, and 20x. These factors cover odd numbers, integers, and decimal, ranging from 10x to pretty

high 20x anisotropy. Comparing with CycleGAN or bicubic interpolation, EMformer has clearer ultrastructure details for any anisotropic scale factors. **b,** After trained on certain anisotropic factors, EMformer can easily perform network transfer learning between various anisotropic data. Here, when EMformer was trained using Cremi data with 10× anisotropy, the EPFL sample can be inferred and restored very well. The SSIM, MSSSIM and PSNR for transfer learned isotropic result are almost the same as those for no transfer learning reconstruction. For the network trained by EPFL data with simulated 8x anisotropy, EMformer can also obtained Cremi structure details almost the same as no transfer learning result. For both case, indicators or ultrastructure details for transfer learning after hyperparameter tuning is just slightly improved than the result without tuning. All scale bars are 500nm.

2.5 EMformer+: session inpainting and isotropic volume reconstruction simultaneously.

At the sample preparation stage, artificial defects such as sample contamination, session wrinkles, and broken, session loss, et. al., always inevitable. For example, sessions after diamond knife cutting may not be collected on the tape which resulting in session missing during imaging; The existence of these phenomena brings further difficulties and challenges to 3D reconstruct of anisotropic vEM data. To solve these problems, we further developed the inpainting function for lost frames and defective frames, which named as EMformer+ and can achieve isotropic reconstruction and session inpainting simultaneously for those defective data with session missing, contamination, blurring, crease and so on.

In the first step, we use self-supervised training data$\{B_i, A_i\}$ consistent with section 2.1, but randomly replace some input frame as blank frame $\{\overline{B}_i\}$ while keep supervision $\{A_i\}$ unchanged. Then perform training of EMformer the same as section 2.1. This process is a simultaneous process of z-axis information estimation and also a x-y plane temporal inpainting for interframe filling. The second step is to remove defective sessions by model inference. Replace the defective area or lost frames of input data with zero to generate $\{\overline{C}_j\}$ or $\{\overline{D}_k\}$, representing validation and testing sets respectively, and form a video clip along the z-axis. The model, only performs inpainting when scale factor s set to 1 and generates $\{C_j'\}$ or $\{D_k'\}$ where contents are filled at blank defective frames based on contextual information. Then replace the defective frames of the original data with the corresponding predictions of the model to generate a complete volume which used as input for isotropic reconstruction inference of the next step. The third step is basically the same as section 2.1 to perform isotropic reconstruction and get $\{\hat{C}_j\}$ or $\{\hat{D}_k\}$. The defect removed data forms video clip along z axis and the model only performs isotropic resolution recovery with scale factor set to anisotropic ratio.

We validated the performance of EMformer+ about single defective session recovery and isotropic reconstruction on EPFL dataset which has ground truth. The original data was simulated degraded as 8x anisotropy and then twelve frames of the anisotropic volume were selected randomly and replaced with blank along z-axis. We train EMformer+ and perform single defective

session recovery and isotropic reconstruction. Surprisingly, the frames that were lost in the simulation were recovered very well, almost have identical ultrastructure and information distribution to the corresponding frames in ground truth (Fig. 5b and Extended Data Fig. 8). Through contextual feature training, the missing frame information is estimated accurately. The SSIM is larger than or near 0.6 and the PSNR is around 25 between the recovery frame and the ground truth for all those lost ones (Supplementary Table 4). This is another embodiment of EMformer's ability to continuously recover spatial information. For isotropic recovery, we compared the results between the result with frame loss and the one without loss in section 2.2 which just degradation but without frame removed. We found that either the ultrastructure or the validation index with ground truth are similar, while SSIM, PSNR, et. al. are just slightly lower than those without frame loss (Fig. 5c and Supplementary Table 5).

Fig. 5: Emformer+ for slice inpainting and isotropic reconstruction simultaneously.

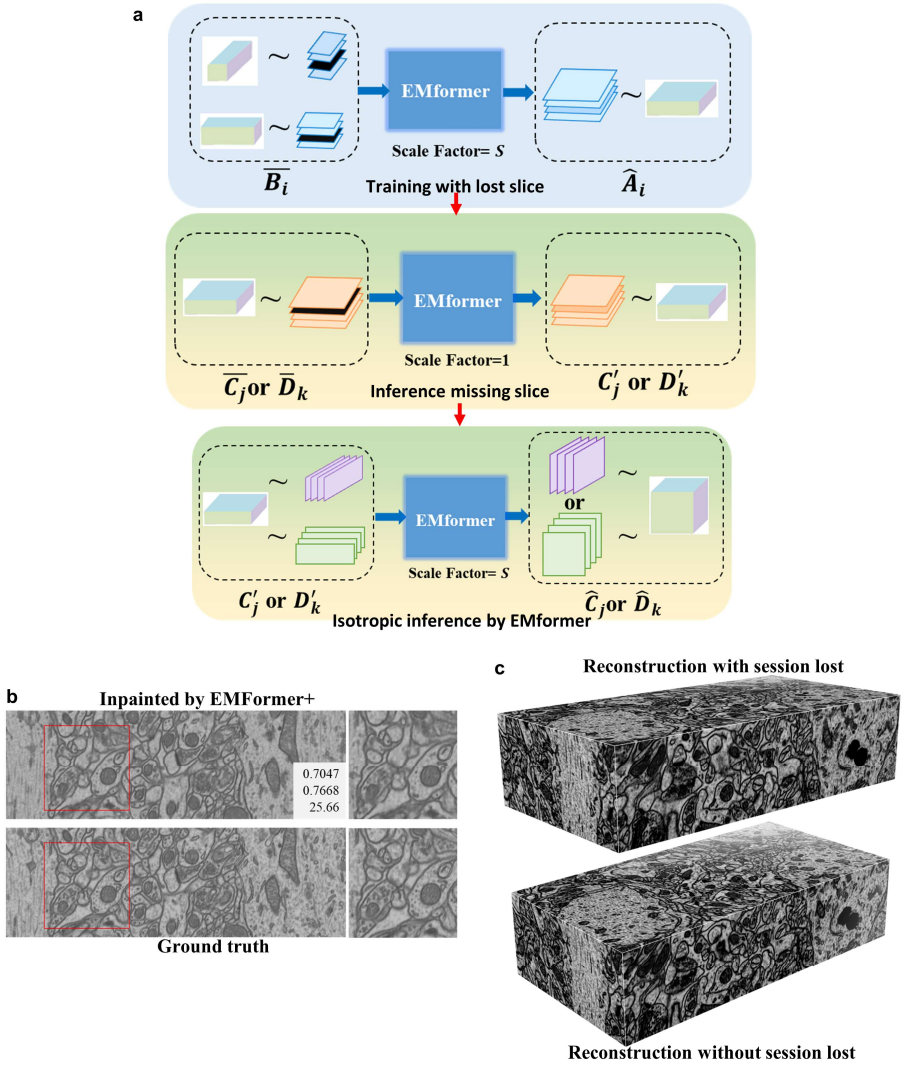


Fig. 5 | Emformer+ for session inpainting and isotropic reconstruction simultaneously. a, Pipeline for EMformer+ to jointly perform single session inpainting and isotropic reconstruction on condition of anisotropic data with defective sessions. $\hat{A}_i$, $\hat{B}_i$, $\hat{C}_j$, $\hat{D}_k$ are the same meaning as Fig. 1b. $\overline{B}_i$ is the result of replacing the random portion of frames in $\hat{B}_i$ with blank, and $\overline{C}_j$, $\overline{D}_k$ are validation set and a test set with single frame missing.

C_j', D_k' are inpainting results of validation set and a test set with frame missing respectively. In the training stage, the model is trained jointly for single frame inpainting and isotropic super-resolution, only changing input data from $\hat{B}_i$ to $\overline{B}_i$ compared to Section 2.1. In the inference process, the model first performs session inpainting and generates $\bar{C}_j, \overline{D}_k$ along the z-axis with anisotropic scale factor of 1, and then performs isotropic reconstruction results C_j', D_k' along x/y axis based on $\bar{C}_j, \overline{D}_k$. **b,** Verifying the model's inpainting performance for lost frames on the EPFL dataset. The 11th session of EPFL data with 8× anisotropy simulation is removed and filled with zero, and then inpainted using EMformer+. The top image is inpainted with EMformer+, while the bottom one is ground truth. **c,** The 3D reconstruction of EMformer+ with session lost (top) and the one by EMformer without session lost (bottom). Both have similar ultrastructure details.

3. Discussion

The isotropic reconstruction method EMformer utilizes the 3D space continuity capability of video transformer to achieve outstanding reconstruction results, and is suitable for the recovery of vEM data with any positive integer or decimal anisotropic factor. It can be applied to FIB-SEM, ssTEM, ssSEM and other different modal data, and enable transferring from one to another even in the case of different anisotropic factors. In addition, the model can handle the case of defective sessions, which improves the robustness against disturbance of real imaging conditions such as contamination、session missing、broken、wrinkles. It provides reliable high-resolution isotropic data for segmentation and ultrastructure analysis, such as neurons, mitochondria, etc.

EMformer can be further optimized and improved. Firstly, the z-axis session degradation is approximated as average pooling in 3D, which can be replaced by other operations such as frame extraction, even replaced by some deep-learning network to simulate the degradation process. Secondly, Emformer achieves section inpainting while isotropic reconstruction. This attachable 2D session recovery function can be further expanded into session denoising、deblurring or combination of them, simply by changing the input data during training accordingly. Thirdly, in custom experiments, many intermediate step of EMformer including multi-scale hierarchy、attention module、plug-and-play up-sampling operator and loss functions can be modified to achieve the best performance in their own experiments.

We have verified that pretrained EMformer on a single dataset exhibit good transferring performance on other modal data. It indicates EMformer pretrained on large-scale vEM data may become a universal vEM reconstruction model without the need for retraining. More test for various dataset will be performed in further study.

In addition, the image inpainting function in EMformer+ will in turn improve the accuracy of image registration. After re-registration, the images are then inpainted and isotropic reconstructed with EMforrmer+, which will further improve the resolution of the result. Therefore, repeated iterations of EMformer+ and image registration will further improve the image quality of vEM reconstruction and bring many potential conveniences to the ultrastructure study of large-scale biological samples.

4. Method

Image Dataset Preparation and Preprocessing

The image datasets used in this work underwent several preparation and preprocessing steps prior to analysis.

Image Datasets Acquisition

Image volumes were obtained from the two sources:

(1) EPFL Dataset: An isotropic FIB-SEM dataset[59] of neural tissue from the hippocampus CA1 region was downloaded from the EPFL website (https://www.epfl.ch/labs/cvlab/data/data-em/). The original isotropic dataset represents a 5x5x5 µm3 volume at 5x5x5 nm/voxel resolution. This was down-sampled by 4x and 8x along z axis to generate simulated anisotropic datasets for validating isotropic reconstruction methods and comparative analysis.

(2) CREMI Dataset: Serial section TEM datasets of neural tissue were obtained from the CREMI repository. The datasets contain 193 sections at 4 x 4 x 40 nm/voxel resolution (Dataset A) (padded version, https://cremi.org/data/).

Image Registration

To align the serial tissue sections along the z-axis, the "Register Virtual Stack Sessions" plugin in ImageJ was utilized. This plugin aligns adjacent image sessions using an automated feature matching algorithm based on SIFT (Scale Invariant Feature Transform)[60-62]. For feature extraction, the “rigid - translate and rotate” model was used. For registration, the alignment models we used were “rigid - translate and rotate” and “moving least squares - maximal warping”. The optimal model was selected based on evaluation of the aligned image stack quality. When needed, elastic transforms were applied to further refine session alignment. This plugin incrementally registers each session to the previous one to generate the aligned 3D image volume, by extracting features between sessions and finding optimal transforms to match them.

Image Cropping

The aligned image stacks were cropped down to specific regions of interest containing relevant ultrastructural tissue details. Cropping served two purposes: 1) focusing the analysis on relevant tissue areas, and 2) removing large black areas introduced by rotation during image alignment. Cropping helped reduce file sizes and excluded unnecessary background regions from further analysis.

Image Enhancement

To enhance image contrast and standardize brightness across the image set, Contrast Limited Adaptive Histogram Equalization (CLAHE) was applied[63]. This technique improves local contrast within small regions while limiting noise amplification. The CLAHE parameters were optimized to avoid over-enhancement artifacts.

Dataset Preparation

The image datasets were extracted and preprocessed into three subsets - training, validation, and test sets. This enables model training, hyperparameter tuning, and unbiased evaluation respectively.

The training dataset comprises the majority of images, used to fit the model parameters. The full training datasets contain several complete blocks of anisotropic EM volumetric data (separation based on the spatial location of the defect sessions). During training, random sub-volumes of a fixed size (e.g. 16x128x128, 32x128x128) are sampled from across these various

blocks to form the training inputs for each batch. This allows the model to train on diverse samples from the broader data distribution. To augment the training data, transformations including rotations and flipping were applied.

A smaller portion of stacks were clip out from the original complete blocks of anisotropic EM volumes for model validation. The validation sub-blocks are disjoint from the training process. The reconstruction of validation data is finished on the best checkpoint for optimal isotropic quality. The validation sub-volumes undergo identical preprocessing as the training data without augmentation.

The test dataset comprises one complete block of anisotropic EM volumes. Since training and inference occur on different dimensions, the same volumes used during training can also serve as test data. Using the same volumes enables direct evaluation on how well the model can super-resolve in the z-dimension based on x-y training. The test volume also undergoes the standard preprocessing similar as training data preparation to eliminate reconstruction artifacts as much as possible.

Network and hyper-parameters design

For the model architecture, we set the stage number of multi-scale transformer as 8, in which the 1st~4th stage, the 4th~7th stage, the 8th stage is for down-sampling, up-sampling and feature fusion respectively. In the first 7 stages, each stage contains 4 (N=4) transformer blocks (TBs), in which 3 TBs contains both MMA and MSA followed by 1 TB contains only MSA. The 8th stage includes 12TBs containing only MSA. The channel sizes for the first 7 stages and the 8th stage are 40 and 60, respectively. For the attention module in TB, we set the number of attention head as 4 for both MSA and MMA. The attention size is set as (2,2,16), (2,2,20) for EPFL and Cremi dataset respectively. Since the larger the attention size, the larger the model size. The model parameters of this two data set are 1.4M, 1.4M respectively.

In the model training stage, we set the training sub-volume size as (16,16,128) and (16,16,160) for EPFL and Cremi data respectively. The batch sizes are all set (2,2,1). We augment the input data by random cropping and 8 orthogonal 3D rotations. The loss function (Equation 3) we use is the weight sum of L_1 (Equation 2) and SSIM loss (Equation 3). The optimizer uses Adam[64]with β1 = 0.9 and β2 = 0.999. The learning rate is set to $1e^{-3}$ and attenuation strategy is not used. It takes about 80k~150k iterations to reach convergence.

$$L_1 = \|I^{SR} - I^{HR}\|_1 \quad (1)$$

$$L_{ssim} = \|ssim(I^{SR}) - ssim(I^{HR})\|_1 \quad (2)$$

$$L_{total} = L_1 + \alpha L_{ssim} \quad (3)$$

where I^{HR} is high resolution image, I^{SR} is super-resolution reconstructed image. The weight parameter α is set to 1 in all our experiments.

In the model inference stage, we generate the testing sub-volumes by cropping the big volume with overlap (8,8,8) in 3D, which size is consistent with training ones. The model inference on those sub-volumes in test-time augmentation manner by 8 orthogonal 3D rotations. The generated isotropic sub-volumes are then stitched in 3D to obtain whole volume according to

cropping coordinates. Linear fusion is adopted for overlapping area during stitching.

The code implementation is under Python 3.8 and PyTorch 1.8.1 cuda version. The training and testing of the model are conducted on a single RTX 3090 NVIDIA GPU card for all experiments. Depending on the size of vEM dataset, it takes about 0.8~1.2 day to train the model, and several hours for model inference.

Evaluation metric

We use four quality assessment metrics to evaluate isotropic reconstruction performance on EPFL simulated degraded data:

(1) SSIM, Structural Similarity Index Measure

SSIM[65] can reflect the perception of human visual system on local structural changes of images. The computation is based on three factors, luminance (l), contrast (c) and structure (s). The overall index is a multiplicative combination of them.

$$\mathrm{SSIM}(x,y) = [l(x,y)]^{\alpha}[c(x,y)]^{\beta}[s(x,y)]^{\gamma} \tag{4}$$

$$l(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}, c(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}, s(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \tag{5}$$

where $\alpha = \beta = \gamma = 1$, $C_3 = C_2/2$. μ_x, μ_y, σ_x, σ_y and σ_{xy} are the local means, standard deviations, and cross-covariance for images x and y respectively. The SSIM values range between 0 to 1, the closer to 1, the better the performance.

(2) MS-SSIM, Multi-Scale Structural Similarity Index Measure

MS-SSIM[66] extends the measurement of SSIM to multiple scales, where the image at scale M is half the image at scale M-1. Through such adaptation, MS-SSIM is more robust than SSIM in face of various resolution images.

$$\mathrm{MSSSIM}(x,y) = [L_M(x,y)]^{\alpha M} \prod_{J=1}^{M} [C_J(x,y)]^{\beta_j} [S_J(x,y)]^{\gamma_j} \tag{6}$$

where $\beta_1 = \gamma_1 = 0.0448$, $\beta_2 = \gamma_2 = 0.2856$, $\beta_3 = \gamma_3 = 0.3001$, $\beta_4 = \gamma_4 = 0.2363$, $\alpha_5 = \beta_5 = \gamma_5 = 0.1333$.

(3) PSNR, Peak Signal to Noise Ratio

PSNR is the ratio between the maximum possible signal power of an image and the power of corrupting noise that affects the image quality. In actual calculation, we estimate the PSNR of an image by comparing it to an ideal clean image with the maximum possible power. High PSNR indicates good image quality.

$$PSNR = 10 \cdot \log_{10}\left(\frac{(L-1)^2}{MSE}\right) = 20 \cdot \log_{10}\left(\frac{L-1}{\sqrt{MSE}}\right) \tag{7}$$

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1} \| x(i,j) - y(i,j) \|^2 \quad (8)$$

where L is the number of maximum possible intensity levels in an image. The x and y are two images to be compared with same size of height n and width m.

(4) LPIPS, Learned Perceptual Image Patch Similarity

LPIPS[67] calculates the perceptual similarity between two images and matches human perception well. LPIPS first generate activations of two 2D images by pre-trained network (such as VGGNet pretrained on ImageNet) and then calculate the feature differences on different layers and spatial dimensions. Since vEM data is 3D, we calculate LPIPS of each frame and average along a certain axis. Low LPIPS score means that image patches are perceptual similar.

$$d(x,y) = \sum_{l} \frac{1}{H_l W_l} \sum_{h,w} \| w_l \odot (\hat{x}^l_{hw} - \hat{y}^l_{hw}) \|_2^2 \quad (9)$$

where $\hat{x}$ and $\hat{y}$ denotes the activation of two image patch. l represents layer of network, H and W represent height and width of the image.

(5) FSC, Fourier Shell Cross-correlation

FSC[68] is a function of spatial frequency, which was first introduced by Harauz and van Heel. It calculates the normalized cross-correlation coefficient between two 3D volumes on corresponding shells with radius r_i in Fourier space. We calculated the FSC between EMformer reconstructed volume and the corresponding GROUND TRUTH volume. The signal-to-noise ratio of vEM data is much better than that of cryoEM data. Therefore, we use the FSC-0.5 standard, which is more suitable for situations where the signal is much more significant than the noise [68].

$$FSC(r_i) = \frac{\sum_{r \in r_i} F_1(r) F_2^*(r)}{\sqrt{\sum_{r \in r_i} F_1^2(r) \cdot \sum_{r \in r_i} F_2^2(r)}} \quad (10)$$

where $F_1(r)$ and $F_2(r)$ are Fourier transform of EMformer reconstructed volume and ground truth volume respectively at radius r in Fourier space; $F_2^*(r)$ is the conjugator of $F_2(r)$.

Evaluation of the Restored vEM volumes in post-process

Reconstruction is performed on the complete datasets using the checkpoints with the best validation performance. The comprehensive evaluation is performed on the restored vEM volumes. The restored output would be imported into MIB for segmentation. GraghCut-based semi-automatic segmentation pipeline and deep-learning based automatic segmentation pipeline were performed in MIB to delineate critical ultrastructural boundaries[69,70].

For the down-sampled EPFL dataset, the neurons randomly selected in ground truth are segmented by labeling foreground and background features using GraphCut tool. We used the totally same pre-defined features to segment corresponding neurons of restored datasets. The segmentation results of these volume restored by various methods and ground truth would be imported into Imaris for surface rendering. We compared the number of disconnected components

and overlapping volume/voxels of surface to assess the pixel similarity and topological relationship similarity between reconstructed data and original data. The F1 score is calculated as the harmonic mean of precision and recall scores of a model. It is defined as follows:

$$F1 - score = \frac{2 \times precision \times recall}{precision + recall} \quad (11)$$

where precision is the number of true positive results divided by the number of all positive results, including those not identified correctly, and recall is the number of true positive results divided by the number of all samples that should have been identified as positive.

For Cremi dataset, we use the segment anything model SAM (https://segment-anything.com/) to assess the topological structural relationship of datasets via segmentation of membrane, vesicle and mitochondria, etc.

Statistical Analysis

Surface rendered models in Imaris were used to characterize neurons and organelle morphologies in 3D. We mainly use Excel to integrate and analyze preliminary statistical data, and GraphPad Prism 9.5 to complete plotting and significance testing analysis.

Data availability

1. EPFL Dataset was downloaded from the EPFL website (https://www.epfl.ch/labs/cvlab/data/data-em/).
2. CREMI Dataset was obtained from the CREMI repository(https://cremi.org/data/)

Code availability

The source codes of EMformer are available at https://github.com/cbmi-group/EMformer.

Reference

1 Terasaki, M. *et al.* Stacked endoplasmic reticulum sheets are connected by helicoidal membrane motifs. *Cell* **154**, 285-296 (2013). https://doi.org/10.1016/j.cell.2013.06.031

2 Muller, A. *et al.* 3D FIB-SEM reconstruction of microtubule-organelle interaction in whole primary mouse beta cells. *J Cell Biol* **220** (2021). https://doi.org/10.1083/jcb.202010039

3 Xu, C. S. *et al.* An open-access volume electron microscopy atlas of whole cells and tissues. *Nature* **599**, 147-151 (2021). https://doi.org/10.1038/s41586-021-03992-4

4 Xu, C. S. *et al.* Publisher Correction: An open-access volume electron microscopy atlas of whole cells and tissues. *Nature* **599**, E5 (2021). https://doi.org/10.1038/s41586-021-04132-8

5 Ryan, K., Lu, Z. & Meinertzhagen, I. A. The CNS connectome of a tadpole larva of Ciona intestinalis (L.) highlights sidedness in the brain of a chordate sibling. *Elife* **5** (2016). https://doi.org/10.7554/eLife.16962

6 Wilson, C. E. *et al.* Taste Bud Connectome: Implications for Taste Information Processing. *J Neurosci* **42**, 804-816 (2022). https://doi.org/10.1523/JNEUROSCI.0838-21.2021

7 Dorkenwald, S. *et al.* Neuronal wiring diagram of an adult brain. *bioRxiv* (2023). https://doi.org/10.1101/2023.06.27.546656

8 Zheng, Z. *et al.* A Complete Electron Microscopy Volume of the Brain of Adult Drosophila melanogaster. *Cell* **174**, 730-743 e722 (2018). https://doi.org/10.1016/j.cell.2018.06.019

9 Peddie, C. J. & Collinson, L. M. Exploring the third dimension: volume electron microscopy comes of age. *Micron* **61**, 9-19 (2014). https://doi.org/10.1016/j.micron.2014.01.009

10 Bosch, C. *et al.* Functional and multiscale 3D structural investigation of brain tissue through correlative in vivo physiology, synchrotron microtomography and volume electron microscopy. *Nat Commun* **13**, 2923 (2022). https://doi.org/10.1038/s41467-022-30199-6

11 Peddie, C. J. *et al.* Volume electron microscopy. *Nat Rev Methods Primers* **2**, 51 (2022). https://doi.org/10.1038/s43586-022-00131-9

12 Morgan, J. L. & Lichtman, J. W. An Individual Interneuron Participates in Many Kinds of Inhibition and Innervates Much of the Mouse Visual Thalamus. *Neuron* **106**, 468-481 e462 (2020). https://doi.org/10.1016/j.neuron.2020.02.001

13 Baena, V. & Terasaki, M. Three-dimensional organization of transzonal projections and other cytoplasmic extensions in the mouse ovarian follicle. *Sci Rep* **9**, 1262 (2019). https://doi.org/10.1038/s41598-018-37766-2

14 Nagai, M. *et al.* Glomerular Cellular Interactions Following Disruption of the Glomerular Basement Membrane in IgA Nephropathy: Ultrastructural Analyses by 3-Dimensional Serial Block-Face Scanning Electron Microscopy. *Kidney Med* **2**, 222-225 (2020). https://doi.org/10.1016/j.xkme.2019.11.003

15 de Senneville, B. D. *et al.* Deciphering tumour tissue organization by 3D electron microscopy and machine learning. *Commun Biol* **4**, 1390 (2021). https://doi.org/10.1038/s42003-021-02919-z

16 Johnson, B. E. *et al.* An omic and multidimensional spatial atlas from serial biopsies of an evolving metastatic breast cancer. *Cell Rep Med* **3**, 100525 (2022). https://doi.org/10.1016/j.xcrm.2022.100525

17 A., E. R. in *Modern Surgical Pathology* (ed Richard J. Cote Noel Weidner, et. al.) 71-84 (2009).

18 Shoemark, A. *et al.* International consensus guideline for reporting transmission electron microscopy results in the diagnosis of primary ciliary dyskinesia (BEAT PCD TEM Criteria). *Eur Respir J* **55** (2020). https://doi.org/10.1183/13993003.00725-2019

19 Shoemark, A. *et al.* Primary ciliary dyskinesia with normal ultrastructure: three-dimensional tomography detects absence of DNAH11. *Eur Respir J* **51** (2018). https://doi.org/10.1183/13993003.01809-2017

20 Takaki, T., Ohno, N., Saitoh, S., Nagai, M. & Joh, K. Podocyte penetration of the glomerular basement membrane to contact on the mesangial cell at the lesion of mesangial interposition in lupus nephritis: a three-dimensional analysis by serial block-face scanning electron microscopy. *Clin Exp Nephrol* **23**, 773-781 (2019). https://doi.org/10.1007/s10157-019-01701-0

21 Schalek, R. *et al.* ATUM-based SEM for High-Speed Large-Volume Biological Reconstructions. *Microscopy and Microanalysis* **18**, 572-573 (2012).

22 Hayworth, K. J. *et al.* Imaging ATUM ultrathin section libraries with WaferMapper: a multi-scale approach to EM reconstruction of neural circuits. *Front Neural Circuits* **8**, 68 (2014). https://doi.org/10.3389/fncir.2014.00068

23 Li, X. *et al.* Large scale three-dimensional reconstruction of an entire Caenorhabditis elegans larva using AutoCUTS-SEM. *J Struct Biol* **200**, 87-96 (2017). https://doi.org/10.1016/j.jsb.2017.09.010

24 Xu, C. S. *et al.* Enhanced FIB-SEM systems for large-volume 3D imaging. *Elife* **6** (2017). https://doi.org/10.7554/eLife.25916

25 Heinrich, H., Bogovic, J. A. & Saalfeld, S. Deep Learning for Isotropic Super-Resolution from Non-isotropic 3D Electron Microscopy. *International Conference on Medical Image Computing and Computer-Assisted Intervention 2017*, 135-143 (2017).

26 Hagita, K., Higuchi, T. & Jinnai, H. Super-resolution for asymmetric resolution of FIB-SEM 3D imaging using AI with deep learning. *Sci Rep* **8**, 5877 (2018). https://doi.org/10.1038/s41598-018-24330-1

27 Deng, S. *et al.* in *International Conference on Medical Image Computing and Computer-Assisted Intervention-MICCAI 2020* Vol. Part of the Lecture Notes in Computer Science book series (LNIP,volume 12265) 11 (2020).

28 Park, H. *et al.* Deep learning enables reference-free isotropic super-resolution for volumetric fluorescence microscopy. *Nat Commun* **13**, 3297 (2022). https://doi.org/10.1038/s41467-022-30949-6

29 Ledig, C. *et al.* Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. *Proc Cvpr Ieee*, 105-114 (2017). https://doi.org/10.1109/Cvpr.2017.19

30 Zhu, J. Y., Park, T., Isola, P. & Efros, A. A. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. *Ieee I Conf Comp Vis*, 2242-2251 (2017). https://doi.org/10.1109/Iccv.2017.244

31 Wang, Z. J. *et al.* STDIN: Spatio-temporal distilled interpolation for electron microscope images. *Neurocomputing* **505**, 188-202 (2022). https://doi.org/10.1016/j.neucom.2022.07.037

32 Pan, M. *et al.* in *Med Image Computing and Computer Assisted Intervention-MICCAI 2023* Vol. Part of the Lecture Notes in Computer Science book series (LNCS,volume 14229) 10 (2023).

33 Lu, C. *et al.* EMDiffuse: a diffusion-based deep learning method augmenting ultrastructural imaging and volume electron microscopy. *bioRxiv, doi：https://doi.org/10.1101/2023.07.12.548636* (2023).

34 Kyungryun Lee, W.-K. J. Reference-Free Isotropic 3D EM Reconstruction using Diffusion Models. *arXiv:2308.01594 [cs.CV]* (2023). https://doi.org/https://doi.org/10.48550/arXiv.2308.01594

35 Xiang, H. Y. *et al.* Deep learning for image inpainting: A survey. *Pattern Recogn* **134** (2023). https://doi.org/https://doi.org/10.1016/j.patcog.2022.109046

36 Zhang, X. B., Zhai, D. H., Li, T. R., Zhou, Y. X. & Lin, Y. Image inpainting based on deep learning: A review. *Inform Fusion* **90**, 74-94 (2023). https://doi.org/10.1016/j.inffus.2022.08.033

37 Pushpalwar, R. T. & Bhandari, S. H. Image Inpainting Approaches - A Review. *Int Conf Adv Compu*, 340-345 (2016). https://doi.org/10.1109/Iacc.2016.70

38 Elharrouss, O., Almaadeed, N., Al-Maadeed, S. & Akbari, Y. Image Inpainting: A Review. *Neural Process Lett* **51**, 2007-2028 (2020). https://doi.org/10.1007/s11063-019-10163-0

39 Rojas, D. J. B., Fernandes, B. J. T. & Fernandes, S. M. M. A Review on Image Inpainting Techniques and Datasets. *Sibgrapi*, 240-247 (2020). https://doi.org/10.1109/Sibgrapi51738.2020.00040

40 Torrado-Carvajal, A. *et al.* Inpainting as a Technique for Estimation of Missing Voxels in Brain Imaging. *Annals of Biomedical Engineering* **49**, 9 (2022). https://doi.org/https://doi.org/10.1007/s10439-020-02556-3

41 Pathak, D., Krähenbühl, P., Donahue, J., Darrell, T. & Efros, A. A. Context Encoders: Feature Learning by Inpainting. *2016 Ieee Conference on Computer Vision and Pattern Recognition (Cvpr)*, 2536-2544 (2016). https://doi.org/10.1109/Cvpr.2016.278

42 Iizuka, S., Simo-Serra, E. & Ishikawa, H. Globally and Locally Consistent Image Completion. *Acm T Graphic* **36** (2017). https://doi.org/https://doi.org/10.1145/3072959.3073659

43 Liu, H. Y. *et al.* PD-GAN: Probabilistic Diverse GAN for Image Inpainting. *2021 Ieee/Cvf Conference on Computer Vision and Pattern Recognition, Cvpr 2021*, 9367-9376 (2021). https://doi.org/10.1109/Cvpr46437.2021.00925

44 Zeng, Y. H., Fu, J. L., Chao, H. Y. & Guo, B. N. Learning Pyramid-Context Encoder Network for High-Quality Image Inpainting. *2019 Ieee/Cvf Conference on Computer Vision and Pattern Recognition (Cvpr 2019)*, 1486-1494 (2019). https://doi.org/10.1109/Cvpr.2019.00158

45 Wan, Z. Y., Zhang, J. B., Chen, D. D. & Liao, J. High-Fidelity Pluralistic Image Completion with

Transformers. *2021 Ieee/Cvf International Conference on Computer Vision (Iccv 2021)*, 4672-4681 (2021). https://doi.org/10.1109/Iccv48922.2021.00465

46 Yu, J. H. *et al.* Free-Form Image Inpainting with Gated Convolution. *2019 Ieee/Cvf International Conference on Computer Vision (Iccv 2019)*, 4470-4479 (2019). https://doi.org/10.1109/Iccv.2019.00457

47 Zheng, C. X., Cham, T. J. & Cai, J. F. Pluralistic Image Completion. *2019 Ieee/Cvf Conference on Computer Vision and Pattern Recognition (Cvpr 2019)*, 1438-1447 (2019). https://doi.org/10.1109/Cvpr.2019.00153

48 Wan, Z. Y. *et al.* Bringing Old Photos Back to Life. *2020 Ieee/Cvf Conference on Computer Vision and Pattern Recognition (Cvpr)*, 2744-2754 (2020). https://doi.org/10.1109/Cvpr42600.2020.00282

49 Nazeri, K., Ng, E., Joseph, T., Qureshi, F. Z. & Ebrahimi, M. EdgeConnect: Structure Guided Image Inpainting using Edge Prediction. *Ieee Int Conf Comp V*, 3265-3274 (2019). https://doi.org/10.1109/Iccvw.2019.00408

50 Ronneberger, O., P., F. & T., B. in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* Vol. Part of the Lecture Notes in Computer Science book series (LNIP,volume 9351) 8 (, 2015).

51 Jingyun Liang, J. C., Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, Luc Van Gool. VRT: A Video Restoration Transformer. *Computer Vision and Pattern Recognition (cs.CV)* (2022). https://doi.org/https://doi.org/10.48550/arXiv.2201.12288

52 Vaswani, A. *et al.* Attention Is All You Need. *Adv Neur In* **30** (2017).

53 Guo, M. H. *et al.* Attention mechanisms in computer vision: A survey. *Comput Vis Media* **8**, 331-368 (2022). https://doi.org/10.1007/s41095-022-0271-y

54 D., L. B. & Kanade, T. in *Proceedings of Imaging Understanding Workshop* 10 (1981).

55 Sun, D. Q., Yang, X. D., Liu, M. Y. & Kautz, J. PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume. *2018 Ieee/Cvf Conference on Computer Vision and Pattern Recognition (Cvpr)*, 8934-8943 (2018). https://doi.org/10.1109/Cvpr.2018.00931

56 Liu, Z. *et al.* Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *2021 Ieee/Cvf International Conference on Computer Vision (Iccv 2021)*, 9992-10002 (2021). https://doi.org/10.1109/Iccv48922.2021.00986

57 Shi, W. *et al.* Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. *arXiv:1609.05158(cs.CV)* **2** (2016).

58 Kirillov A. *et al.* Segment Anything. *arXiv preprint arXiv:2304.02643* (2023).

59 Lucchi, A., Li, Y. & Fua, P. in *2013 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, Portland, OR, USA, 2013).

60 Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nature Methods* **9**, 676-682 (2012). https://doi.org/10.1038/nmeth.2019

61 Lowe, D. G. in *International Conference on Computer Vision* (1999).

62 Lowe, D. G. Distinctive image features from scale-invariant keypoints. *Int J Comput Vision* **60**, 91-110 (2004). https://doi.org/Doi 10.1023/B:Visi.0000029664.99615.94

63 Zuiderveld, K. in *Graphics gems IV* 12 (1994).

64 Kingma, D. P. & Ba, J. in *the 3rd International Conference for Learning Representations* (San Diego, 2015).

65 Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE Trans Image Process* **13**, 600-612 (2004). https://doi.org/10.1109/tip.2003.819861

66 Wang Z., Simoncelli E. P. & C., B. A. in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003* (IEEE, Pacific Grove, CA, USA, 2003).

67 Zhang, R., Isola, P., Efros, A. A., Shechtman, E. & Wang, O. in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, Salt Lake City, UT, USA, 2018).

68 van Heel, M. & Schatz, M. Fourier shell correlation threshold criteria. *J Struct Biol* **151**, 250-262 (2005). https://doi.org/10.1016/j.jsb.2005.05.009

69 Belevich, I., Joensuu, M., Kumar, D., Vihinen, H. & Jokitalo, E. Microscopy Image Browser: A Platform for Segmentation and Analysis of Multidimensional Datasets. *PLoS Biol* **14**, e1002340 (2016). https://doi.org/10.1371/journal.pbio.1002340

70 Belevich, I. & Jokitalo, E. DeepMIB: User-friendly and open-source software for training of deep learning network for biological image segmentation. *PLoS Comput Biol* **17**, e1008374 (2021). https://doi.org/10.1371/journal.pcbi.1008374

Acknowledgments

The work was supported by the National Natural Science Foundation of China (322371248 to Y. Z., 31971289, 91954201 to G. Y., 92254306 and 31925026 to F. S.), the National Key Research and Development Program of China (2021YFF0704300 to Y. Z. and 2021YFA1301500 to F. S.), the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB37040402 to G.Y.). We thank Y. Feng and C. L. Liu from the Center for Biological Imaging (CBI), Institute of Biophysics, Chinese Academy of Science, for their help in image segmentation. We acknowledge the services provided by X. X. Li from CBI, Institute of Biophysics, Chinese Academy of Science for ssSEM data collecting. We thank S. D. Yang from Institute of Biophysics, Chinese Academy of Science for his help of data pre-processing, and L. R. Li from Institute of Automation, Chinese Academy of Science for her help of interpolation method survey.

Author contributions

F. S. conceived the concept. Y. Z. supervised the study with the support by F. S. and G. Y.. J. H. proposed the model architecture, conducted all model training and testing. J. H. conducted code replication of supervised method and CycleGAN. W. H. S. performed pre-processing of test data and segmentation of EPFL data. J. H. performed segmentation of Cremi data. Y. Z. analyzed all data and result. Y. Z. wrote the manuscript with input from all coauthors. Y. Z., G. Y. and F. S. acquired funding for this project.