# FIB-SEMBENCH: A BENCHMARK FOR ISOTROPIC SUPER-RESOLUTION OF FIB-SEM VOLUMES

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Volumetric electron microscopy, particularly focused ion beam scanning electron microscopy (FIB-SEM), enables three-dimensional visualization of cellular ultrastructure, but acquiring high-resolution isotropic volumes is constrained by high cost and low throughput. 3D super-resolution (SR) offers a way to recover details from low-resolution volumes, while existing SR evaluations rely mainly on image-fidelity metrics. Whether higher image-level fidelity reflects faithful 3D structure or accurate mitochondrial volume estimates remains unclear. To examine this issue, we present FIB-SEMBench, a standardized benchmark for $2\times$ isotropic SR ($8 \rightarrow 4$ nm). The benchmark covers four biologically diverse datasets spanning mouse and *Drosophila* tissues, with a physics-inspired degradation pipeline and spatially separated data partitioning. We benchmark four 3D SR approaches in terms of image-level fidelity, 3D structural preservation, mitochondrial volume accuracy, and computational efficiency. Our evaluation reveals that higher image-level fidelity does not necessarily correspond to better 3D structure preservation or more accurate mitochondrial volume estimates.

## 1 INTRODUCTION

Volumetric electron microscopy (vEM) provides nanometer-scale 3D imaging of cellular ultrastructure, with FIB-SEM enabling high-resolution volumetric imaging over extended biological volumes (Xu et al., 2021). High-resolution isotropic imaging (e.g., 4 nm) preserves fine structural details required for quantitative organelle analysis; however, achieving this resolution entails longer acquisition times and increased costs. FIB-SEM datasets are commonly acquired at isotropic voxel sizes around 8 nm (Scheffer et al., 2020). Compared with 4 nm sampling, 8 nm sampling resolves fewer fine ultrastructural details (Xu et al., 2021). This trade-off makes SR an attractive strategy for recovering high-resolution details from lower-resolution volumes.

Deep learning-based SR has been applied to electron microscopy. Existing studies have primarily focused on either 2D slice-wise restoration (Qian et al., 2020; Suveer et al., 2019) or anisotropic-to-isotropic reconstruction (Qiu et al., 2026; Longmi & Pan, 2026). These approaches differ from 3D isotropic SR, where the input volume is already isotropic and the objective is to increase spatial resolution consistently across all three dimensions. This distinction motivates a systematic examination of how different deep learning paradigms perform in this isotropic setting.

Evaluating these methods, however, requires considering how reconstruction quality should be assessed. Standard metrics such as PSNR and SSIM are widely used to measure reconstruction fidelity, but their validity in medical image quality analysis has been questioned (Breger et al., 2025). In EM segmentation, prior work has shown that conventional metrics do not fully capture the downstream impact of these errors (Plaza & Funke, 2018). Whether image-level fidelity corresponds to preservation of 3D structures or the accuracy of mitochondrial volume estimates in isotropic FIB-SEM SR remains uncharacterized.

To examine this question, we introduce FIB-SEMBench, a standardized benchmark for evaluating controlled $2\times$ isotropic SR of FIB-SEM volumes (Figure 1). Built upon diverse 4 nm isotropic datasets from OpenOrganelle, the benchmark incorporates a physics-inspired degradation pipeline to compare representative SR paradigms, including volumetric regression models and diffusion models, while interpolation and slice-wise methods serve as reference baselines. FIB-SEMBench evalu-



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

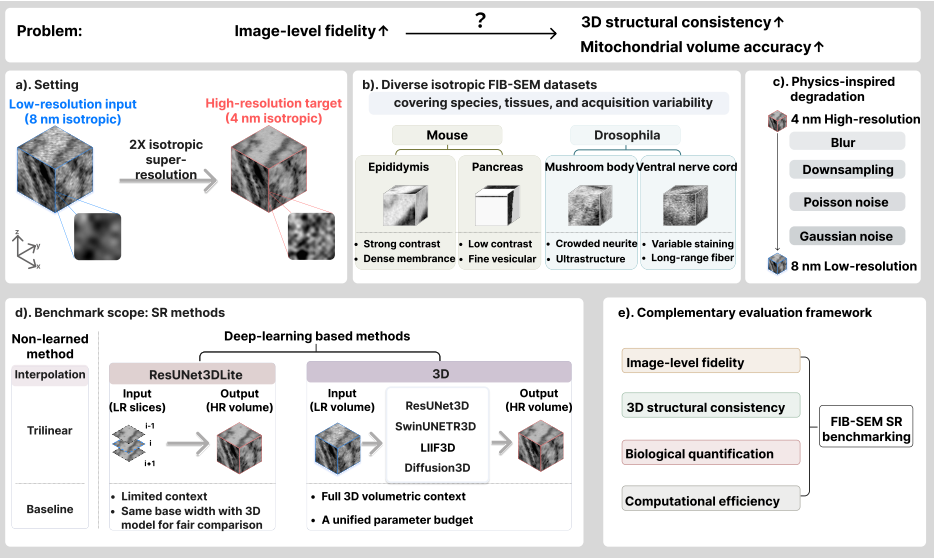


Figure 1: Overview of FIB-SEMBench. a) Supervised reconstruction from low-resolution isotropic inputs to high-resolution isotropic targets. b) Four isotropic FIB-SEM datasets with diverse biological specimens and acquisition protocols. c) Physics-inspired degradation pipeline. d) Comparison of interpolation baseline, ResUNet3DLite, and 3D SR methods. e) Complementary suite of evaluation.

ates 3D SR across four criteria: image-level fidelity, 3D structural consistency, biological quantification, and computational efficiency under standardized settings.

The main contributions of this work are summarized as follows:

- We establish FIB-SEMBench, a standardized benchmark for controlled $2\times$ isotropic FIB-SEM SR, comprising biologically diverse 4 nm datasets, a physics-inspired degradation pipeline, and unified training-evaluation protocols.
- We find that image-level fidelity does not fully capture 3D structural quality, as shown by differences in cross-plane structural continuity and structured reconstruction errors across SR models.
- We demonstrate that image-level fidelity does not reliably predict HR-referenced mitochondrial volume accuracy in FIB-SEM SR.

## 2 Related Works

**Volumetric SR methods** Volumetric SR draws on developments in both image SR and 3D biomedical imaging. Image SR has progressed from CNN baselines (Dong et al., 2015; Lim et al., 2017; Zhang et al., 2018b) to Swin Transformers (Liang et al., 2021), continuous coordinate representations (Chen et al., 2021), and diffusion models (Saharia et al., 2023). 3D biomedical imaging has established convolutional (Çiçek et al., 2016; Milletari et al., 2016) and Transformer-based models (Hatamizadeh et al., 2022; 2021) for modeling 3D spatial context. Many vEM SR studies have explored reconstructing isotropic volumes from anisotropic acquisitions (Yang et al., 2026; Khateri et al., 2023; Qiu et al., 2026; Chen et al., 2026). These approaches offer a rich design space for 3D SR, but the isotropic-to-isotropic SR setting in FIB-SEM lacks established baselines.

**Image-level fidelity assessment** SR evaluation relies on image-level metrics such as MSE (despite known limitations (Wang & Bovik, 2009)), SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018a). FSC (Van Heel & Schatz, 2005) and its variants including modified FSC (Penczek, 2020), FSC-AUC (Jeon et al., 2024), and localized FSC measures (Verbeke et al., 2024; Ramírez-Aportela et al., 2021) provide frequency-domain measures of 3D reconstruction quality.

**Task-aware evaluation in electron microscopy** Task-aware evaluation considers whether reconstructed images preserve information relevant to downstream biological analysis or clinical analysis. Recent studies have evaluated reconstruction methods using both image-level metrics and downstream task performance in biomedical imaging (Amoros et al., 2025) and electron microscopy (Yang et al., 2026; Du et al., 2026). However, these studies primarily assess whether a reconstruction method benefits downstream analysis. For FIB-SEM SR, the relationship between image-level fidelity and downstream biological utility remains insufficiently characterized.

# 3 METHODS

## 3.1 PROBLEM FORMULATION AND BENCHMARK OVERVIEW

We formulate FIB-SEM SR as the reconstruction of a high-resolution isotropic volume from a low-resolution isotropic input. Given a 4 nm isotropic reference volume $v^{\mathrm{HR}}$, we generate the corresponding 8 nm isotropic input using a controlled degradation operator $\mathcal{D}_{\boldsymbol{\psi}}$:

$$v^{\mathrm{LR}} = \mathcal{D}_{\boldsymbol{\psi}}(v^{\mathrm{HR}}). \tag{1}$$

The SR model $f_\theta$ reconstructs

$$\hat{v}^{\mathrm{HR}} = f_\theta(v^{\mathrm{LR}}). \tag{2}$$

We examine whether higher image-level fidelity consistently corresponds to faithful 3D structure and reliable downstream biological quantification. Each reconstruction is evaluated from three perspectives:

$$\mathcal{E}\left(\hat{v}^{\mathrm{HR}}, v^{\mathrm{HR}}\right) = \left(S_{\mathrm{img}}, S_{\mathrm{vol}}, S_{\mathrm{bio}}\right), \tag{3}$$

where $S_{\mathrm{img}}$ measures image-level fidelity, $S_{\mathrm{vol}}$ assesses 3D structural preservation, and $S_{\mathrm{bio}}$ measures downstream biological utility through mitochondrial volume quantification.

This benchmark was designed around four methodological principles: leakage-free spatial partitioning, informative and spatially diverse ROI selection, controlled LR generation, and a consistent evaluation protocol across SR architectures using both image-level and biological criteria.

## 3.2 BENCHMARK CONSTRUCTION

### 3.2.1 DATASET SELECTION AND RESOLUTION VALIDATION

We selected four datasets from the OpenOrganelle atlas (Xu et al., 2021) to capture biological and technical variability. These datasets span distinct species, tissue types, and acquisition processes (Table A1 in Appendix A). Such variations influence electron-sample interactions, leading to inherent differences in signal-to-noise ratio, contrast, and potential imaging artifacts across datasets. Each dataset provides an isotropic volume. The full-resolution level ($S_0$) has a voxel size of 4 nm, and the corresponding pyramid level ($S_1$) has a voxel size of 8 nm. To verify the consistency between $S_0$ and $S_1$, we compared the released $S_1$ with synthetically downsampled versions of $S_0$. Agreement was quantified using mean squared error (MSE) and structural similarity index (SSIM). Details are shown in Table A1 in Appendix A.

### 3.2.2 LEAKAGE-FREE SPATIAL PARTITIONING

We partitioned each volume along the $z$-axis into training, validation, and test regions ($70\%/15\%/15\%$, respectively). To prevent spatial data leakage, we introduced an exclusion margin equal to the axial depth of the input patch ($L_z = 64$ slices), discarding any patch that extends into this margin. Consequently, every retained input patch was fully contained within a single partition.

### 3.2.3 INFORMATIVE AND SPATIALLY DIVERSE ROI SELECTION

Whole FIB-SEM volumes suffer from class imbalance (Figure A1 in Appendix A), making random sampling biased toward frequent structures and less likely to capture rare targets. We identified informative candidate regions on the validated $S_1$ volume using patch-level image features, including fill proportion, and post-smoothed intensity variation. We retained the top $M = 10{,}000$ candidates

and applied Farthest Point Sampling (FPS) to their $S_0$ spatial coordinates to reduce spatial redundancy. Candidate coordinates were identified at the $S_1$ level and mapped to the corresponding $S_0$ coordinates through a shared reference system (Equation A1 in Appendix A).

### 3.3 Controlled Physics-inspired Degradation Model

Training pairs are often generated by downsampling alone, creating an idealized correspondence between HR and LR volumes that does not fully reflect degradation present in practical imaging. We adopt a controlled physics-inspired degradation model to synthesize training pairs. HR volumes were first linearly normalized to the range [-1,1]. The degradation process was defined as:

$$y = \mathcal{D}_\phi(v) = C \circ N_g \circ N_p \circ S \circ B(v), \tag{4}$$

where $v$ denotes the normalized HR volume and $y$ the synthesized LR volume. $B$, $S$, $N_p$, and $N_g$ represent spatial blur, $2\times$ downsampling, Poisson noise, and Gaussian noise, respectively, while $C$ clips intensities to the valid range. These operators capture three degradation factors: blur, resolution reduction, and intensity noise (Weigert et al., 2018). All degradation parameters were fixed across experiments; see Table A2 in Appendix B.

### 3.4 Model architectures and adaptation

**Model selection and common protocol** We evaluated four distinct reconstruction paradigms, including convolutional, transformer-based, implicit neural representation, and diffusion-based approaches. The corresponding evaluated models were ResUNet3D, SwinUNETR3D, LIIF3D, and Diffusion3D. These models were derived from 3D U-Net (Çiçek et al., 2016), Swin UNETR (Tang et al., 2022), LIIF (Chen et al., 2021), and SR3 (Saharia et al., 2023), respectively, and adapted to the common 3D isotropic $2\times$ SR task while retaining their core reconstruction formulations. The model sources and principal task-specific adaptations are summarized in Table A3 in Appendix B. All models followed the same dataset splits, preprocessing procedures, degradation process, and training patch size, with the optimization budget also kept consistent across models.

**Model-specific adaptations** For the CNN- and transformer-based models, residual learning was used: the network predicted an HR correction that was added to the trilinearly upsampled input to obtain the final SR output. ResUNet3D used normalization-free residual blocks to accommodate the small batch sizes required for 3D training. The final convolution of the residual branch was zero-initialized. SwinUNETR3D incorporated a reconstruction head comprising 3D convolutions and pixel shuffle to transform LR transformer features into an HR residual. LIIF3D extended implicit coordinate-based reconstruction to volumetric data through 3D feature unfolding and 3D coordinate encoding. Diffusion3D formulated 3D SR as conditional diffusion using a packed representation. The model used the LR input and the interpolation baselines as conditioning variables and was trained with velocity prediction.

**Training and checkpoint selection** During training, we periodically evaluated each model on the validation set using PSNR and SSIM. We selected the checkpoint with the highest validation PSNR for final evaluation on the held-out test set. Details are provided in Table A4 in Appendix B.

### 3.5 Evaluation protocol

**Image-level fidelity** Image-level fidelity was quantified by PSNR and SSIM to measure intensity fidelity and structural similarity between reconstructed volumes and the corresponding HR reference volumes. Frequency-domain consistency was assessed by FSC-AUC, computed from Fourier shell correlation curves, while LPIPS was used to quantify perceptual similarity. All metrics were computed under standardized preprocessing conditions.

**3D structural consistency** We assessed 3D structural consistency from two complementary perspectives. First, matched regions from ResUNet3D and ResUNet3DLite were compared across the three orthogonal orientations (XY, XZ, and YZ), with paired differences computed over corresponding held-out test regions. Second, structured reconstruction artifacts were characterized using orientation-dependent residual analysis and stride-locked phase-bias analysis. For a given spatial

period $P$, residuals were grouped by spatial phase $(y \bmod P, x \bmod P)$, and the phase-bias score was defined as the standard deviation of the phase-specific mean residuals after subtracting their overall mean.

**Biological quantification** To evaluate downstream impact, HR references and reconstructed volumes were segmented using MitoNet model (Conrad & Narayan, 2023) (Empanada framework) under an identical 3D instance segmentation pipeline. We quantified mitochondrial volume preservation using absolute relative volume error between the SR and HR foreground voxel counts, while segmentation consistency was measured using Dice and IoU. The volume error is defined in Equation A2 in Appendix B.

**Computational efficiency** We evaluated trainable parameters, peak GPU memory, and inference time for a single $64^3$ input patch under the same GPU setting. Peak memory was measured during a forward pass, while inference time was measured over 50 runs after 10 warm-up runs, using GPU synchronization for timing; the reported time was the mean of the fastest 20% of runs.

## 4 RESULTS

### 4.1 IMAGE-LEVEL FIDELITY VARIES ACROSS SR MODELS AND METRICS

Regression-based SR models, including ResUNet3D, SwinUNETR3D, and LIIF3D, consistently outperformed trilinear interpolation in terms of PSNR, SSIM, and FSC-AUC across all four datasets (Figure 2). By comparison, Diffusion3D showed dataset-dependent performance. It exceeded trilinear interpolation across all three fidelity metrics on the epididymis dataset, improved PSNR and SSIM but not FSC-AUC on the pancreas dataset, and underperformed trilinear interpolation on both *D*rosophila datasets (see Table A5 in Appendix C for its details).

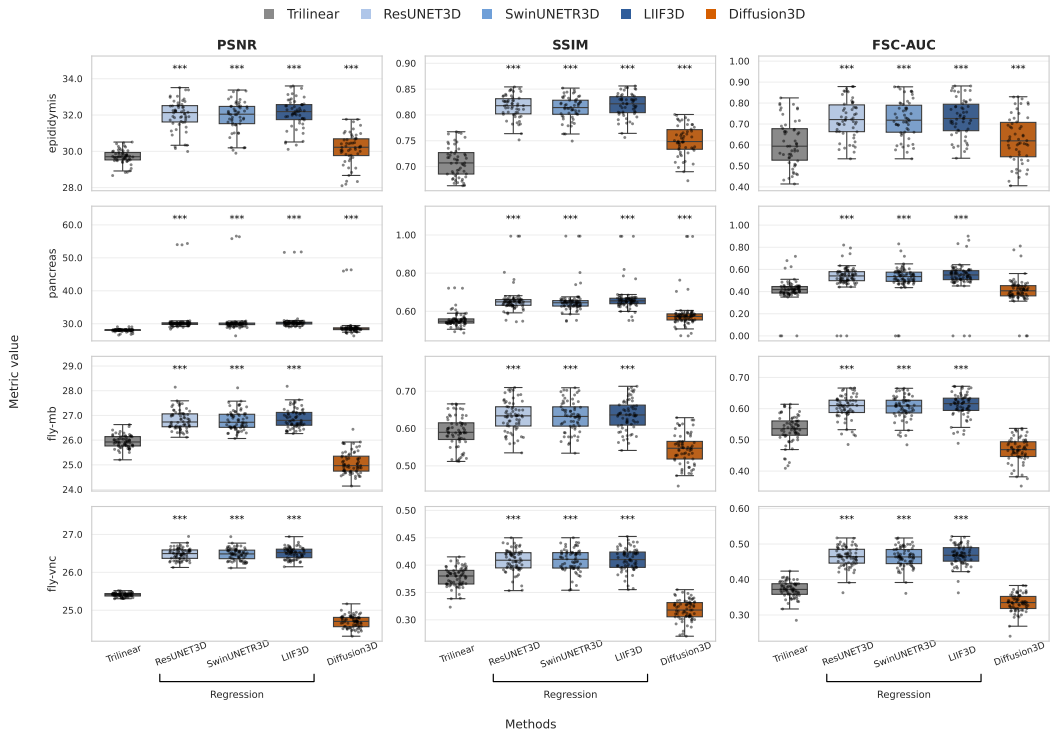


Figure 2: Image-level fidelity across SR architectures.

Despite such variable fidelity performance, Diffusion3D achieved the lowest LPIPS on all four datasets. This contrast highlights that model rankings depend on the choice of metric, as illustrated

in Figure 3. For the regression-based models, Kendall's W indicated that rankings were largely consistent across samples (Table A6 in Appendix C).

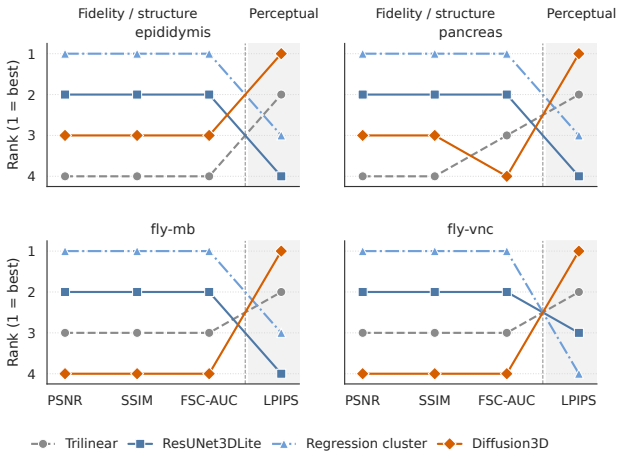


Figure 3: Rank of compared reconstruction methods across fidelity and perceptual metrics.

## 4.2 Image-level fidelity does not fully capture 3D reconstruction quality

### 4.2.1 3D context improves cross-plane structural continuity

To examine the contribution of volumetric context, we evaluated ResUNet3D and ResUNet3DLite under the same residual-learning framework and input/output settings. ResUNet3D processes the interpolated volume with 3D convolutions, whereas ResUNet3DLite processes each XY slice independently with 2D convolutions, so only ResUNet3D learns residual features jointly across adjacent slices. In the representative matched regions of the four datasets, both variants preserved the overall structural organization in the XY views. Differences were more apparent in the XZ and YZ views, where the Lite reconstruction exhibited horizontal streaking, local stretching, and less consistent boundary continuity than ResUNet3D (Figure 4). Across all four datasets, these qualitative observations were consistent with the quantitative comparison, with ResUNet3D showing higher mean values and positive differences across patches (Table A7 in Appendix D; Figure A2).

### 4.2.2 Structured and periodic reconstruction errors

To examine what may underlie the discrepancy between image-level fidelity and volumetric structure, we analyzed the spatial organization of reconstruction errors. Across signed residual maps, local RMS error, and DoG-filtered residuals, reconstruction errors exhibited spatially structured patterns (Figure 5). Similar patterns were observed across the coronal and sagittal views (Figures A3 and A4 in Appendix D).

We then examined whether these spatial patterns showed systematic periodic organization using phase-dependent reconstruction errors across different stride periods. As shown in Figure 6, phase bias remained weak and spatially diffuse at $P = 2$, became more distinct at $P = 4$, and showed stronger and more spatially organized patterns at $P = 8$ across the compared learning-based methods. This progressive phase-dependent organization indicates stride-locked periodic residual patterns. Such spatially localized and phase-dependent errors may be only weakly reflected in image-level fidelity scores, despite representing systematic reconstruction deviations.

## 4.3 Image-level quality does not reliably predict mitochondrial volume accuracy

To examine whether image-level quality differences were associated with downstream quantitative measurements, we compared four representative 3D SR architectures (ResUNet3D, SwinUNETR3D, LIIF3D, and Diffusion3D) across 73 QC-qualified samples with matched HR reference

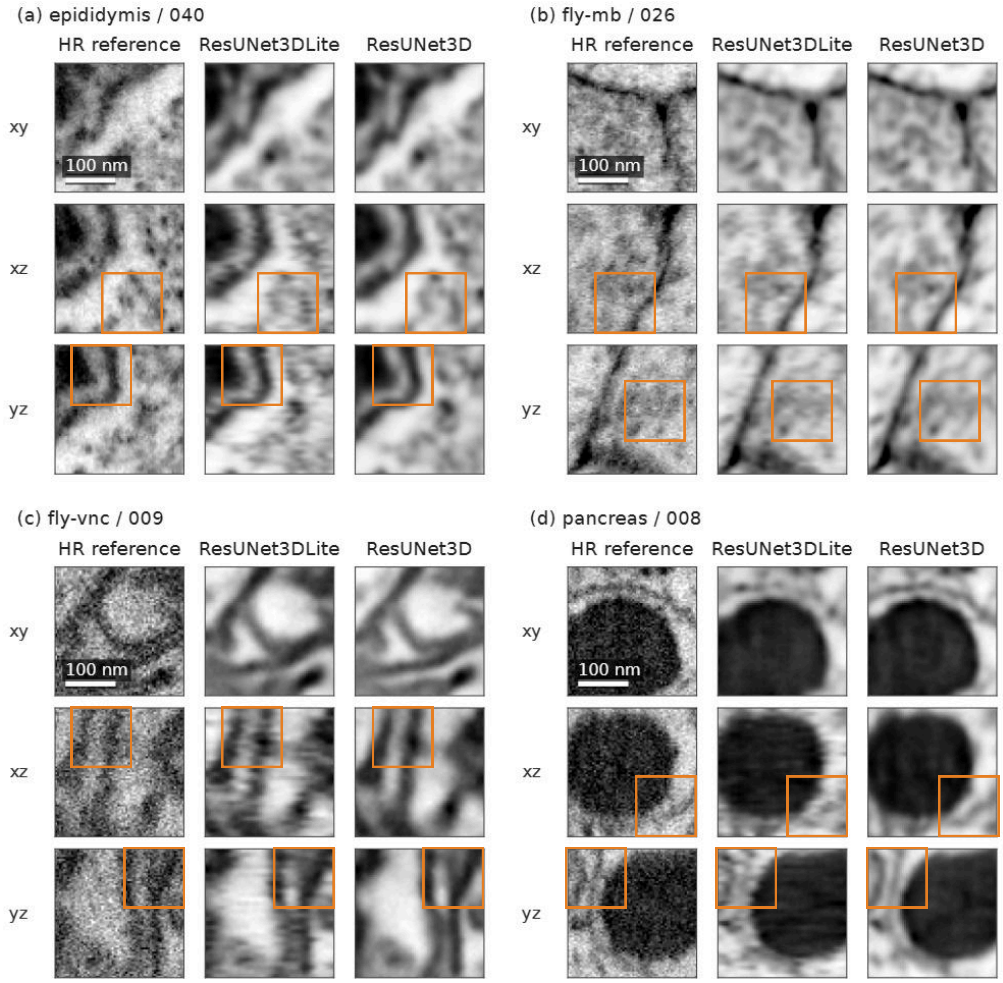


Figure 4: Orthogonal-view comparison of ResUNet3D and ResUNet3DLite.

(15 epididymis, 28 pancreas, 15 fly-mb, and 15 fly-vnc). Figure 7(a) shows the relationships between absolute relative volume error and four image-level fidelity metrics across the 292 method–sample comparisons. PSNR, SSIM, and FSC-AUC showed weak association with volume error, whereas LPIPS showed a positive association ($\rho = 0.310$), indicating that larger perceptual distances tended to coincide with larger volume errors. These results show that image-level fidelity metrics did not consistently correspond to more accurate HR-referenced mitochondrial volume estimates.

Because Dice similarity and foreground-volume error were calculated from the same HR/SR mask pairs, their association was interpreted as an internal consistency analysis. Dice and IoU values for all reconstruction methods are reported (Figure A5 and FigureA6 in Appendix E). Across all 292 method–sample pairs, Dice was negatively correlated with absolute relative foreground-volume error (Spearman $\rho = -0.581$, $p = 1.0 \times 10^{-27}$). Lower HR-referenced segmentation overlap generally coincided with larger deviations in the segmentation-derived volume proxy.

The negative association was also present within each reconstruction architecture: $\rho = -0.570$ for LIIF3D, $\rho = -0.551$ for ResUNet3D, $\rho = -0.538$ for SwinUNETR3D, and $\rho = -0.651$ for Diffusion3D (all $p < 10^{-6}$). Its strength varied by dataset. Strong associations were observed in epididymis ($\rho = -0.766$, $p = 9.6 \times 10^{-13}$), fly-mb ($\rho = -0.749$, $p = 6.0 \times 10^{-12}$), and pancreas ($\rho = -0.599$, $p = 3.0 \times 10^{-12}$). In fly-vnc, the association was weaker and did not reach statistical significance ($\rho = -0.226$, $p = 0.083$). Reduced segmentation consistency was associated with larger foreground-volume errors, though with varying magnitudes across datasets.

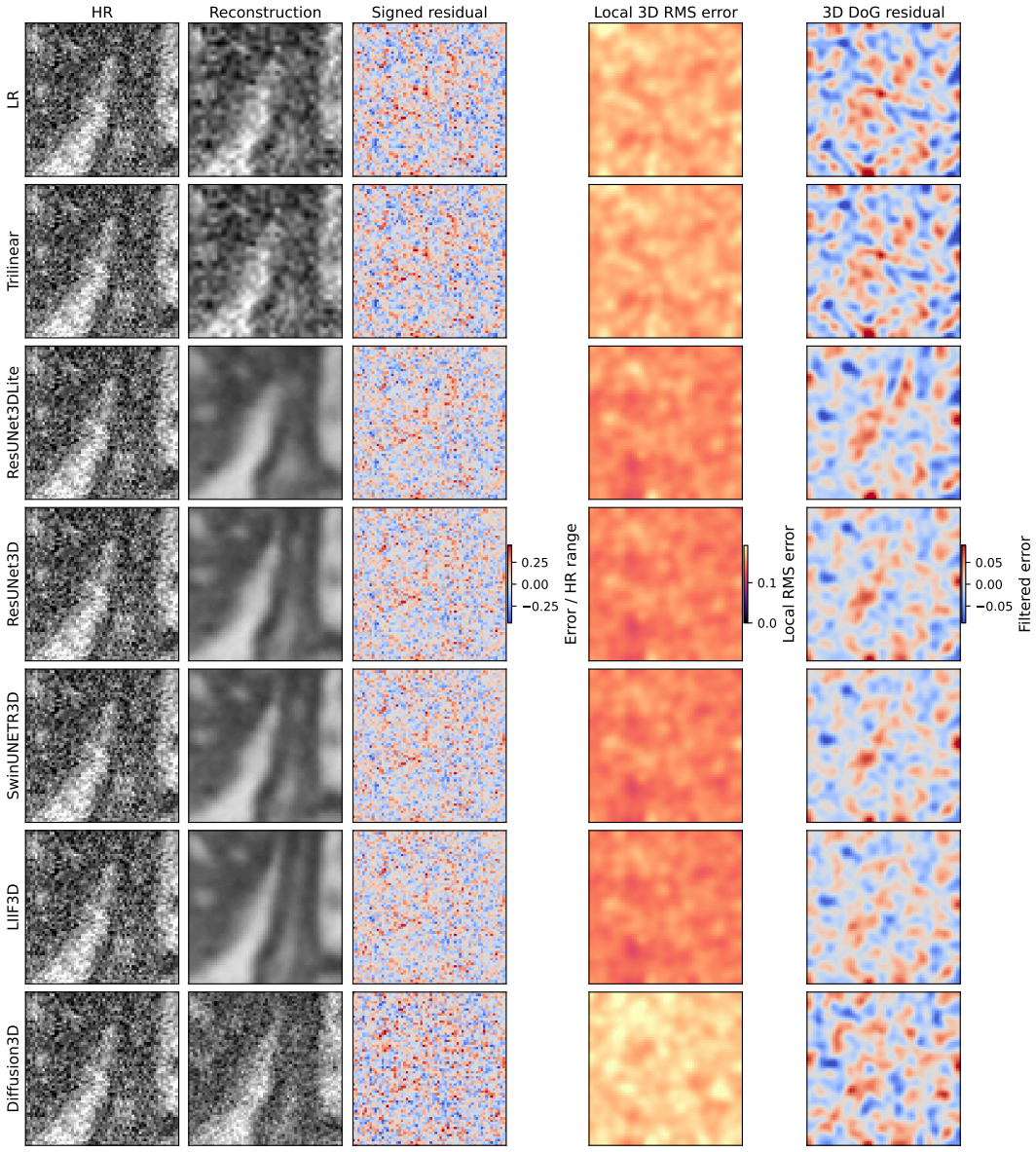


Figure 5: Representative analysis of orientation-dependent reconstruction residuals in the fly-vnc dataset.

### 4.4 Reconstruction paradigms show different computational costs

Despite similar parameter counts, the four 3D SR architectures differed substantially in inference time and peak GPU memory. Inference time ranged from 0.013 s per volume with ResUNet3D to 3.220 s with LIIF3D, while peak GPU memory ranged from 213.4 MB to 3664.7 MB (Table A9 in Appendix F). ResUNet3D achieved a median absolute relative volume error of 79.92%, whereas Diffusion3D had a lower median error of 50.15% but at the cost of 1.773 s per volume. SwinUNETR3D and LIIF3D had median volume errors of 100.00% and 93.79%, respectively, with inference times of 0.031 s and 3.220 s. ResUNet3D and Diffusion3D lay on the Pareto frontier, with ResUNet3D offering faster inference and Diffusion3D achieving lower median volume error (Figure 7(b)).

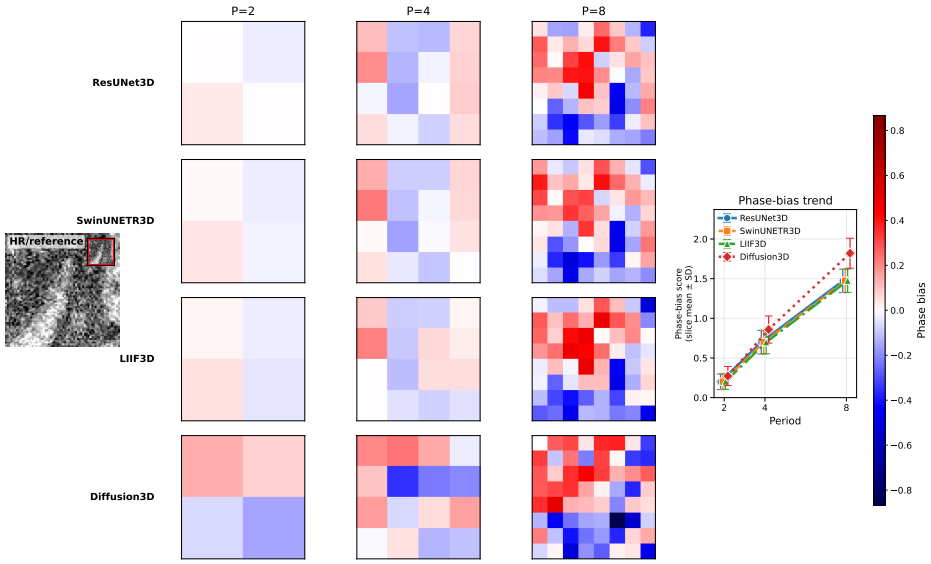


Figure 6: Stride-locked phase artifact analysis across representative 3D reconstruction models. Left panels show normalized phase-bias maps, where red and blue indicate positive and negative phase bias, respectively. Right panels summarize the corresponding phase-bias scores as a function of stride period ($P$).

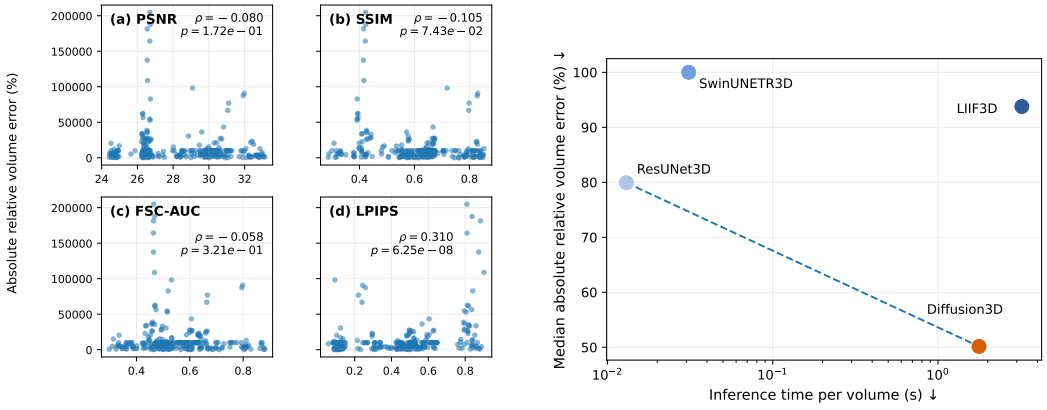


Figure 7: (a) Relationships between absolute relative volume error and four image-level fidelity metrics across matched method–sample comparisons. (b) Trade-off between inference time per volume and median absolute relative volume error across 3D SR architectures.

## 5 Conclusion

We established FIB-SEMBench, a standardized benchmark for controlled 2× isotropic FIB-SEM SR, to examine how well image-level fidelity reflects 3D structural and biological quality. Our results show that image-level fidelity does not fully capture 3D structural quality or reliably predict HR-referenced mitochondrial volume accuracy, indicating that image-level fidelity provides useful but incomplete evidence of 3D reconstruction quality. These findings highlight the need to evaluate FIB-SEM SR beyond image-level fidelity, with complementary assessment of 3D structural quality and biological quantification.

### AI USE STATEMENT

In this work, we used generative AI tools to assist with feedback on experimental design and the interpretation and presentation of experimental results. We additionally used generative AI tools for drafting and editing parts of the manuscript, improving writing clarity and organization, and literature search, identifying relevant references, and formatting citations. All AI-assisted technical statements were manually reviewed and verified by the authors. All reported quantitative results were obtained from author-run experiments and were not generated or modified by generative AI. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

### ETHICS STATEMENT

### REPRODUCIBILITY STATEMENT

All benchmark stages followed fixed configurations and random seeds. Code and detailed configuration settings are provided as supplementary material and will be released publicly upon publication.

### AUTHOR CONTRIBUTIONS

### ACKNOWLEDGMENTS

## REFERENCES

Mario Amoros, Manuel Curado, and Jose F. Vicent. Evaluating super-resolution models in biomedical imaging: Applications and performance in segmentation and classification. *Journal of Imaging*, 11(4), 2025. ISSN 2313-433X. doi: 10.3390/jimaging11040104. URL https://www.mdpi.com/2313-433X/11/4/104.

Anna Breger, Ander Biguri, Malena Sabaté Landman, Ian Selby, Nicole Amberg, Elisabeth Brunner, Janek Gröhl, Sepideh Hatamikia, Clemens Karner, Lipeng Ning, Sören Dittmer, Michael Roberts, AIX-COVNET Collaboration, and Carola-Bibiane Schönlieb. A Study of Why We Need to Reassess Full Reference Image Quality Assessment with Medical Images. *Journal of Imaging Informatics in Medicine*, 38(6):3444–3469, March 2025. ISSN 2948-2933. doi: 10.1007/s10278-025-01462-1. URL https://link.springer.com/10.1007/s10278-025-01462-1.

Bohao Chen, Yanchao Zhang, Yanan Lv, Chenxun Deng, Hua Han, and Xi Chen. Skelem: Training-signal decoupling of skeleton and diffusion for self-supervised axial super-resolution in volume microscopy. In *Computer Vision – ECCV 2026*, Lecture Notes in Computer Science. Springer Nature Switzerland, 2026.

Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8628–8638, 2021.

Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pp. 424–432. Springer, 2016.

Rowan Conrad and Kedar Narayan. Instance segmentation of mitochondria in electron microscopy images with a generalist deep learning model trained on a diverse dataset. *Cell Systems*, 14(1): 58–71.e5, 2023. doi: 10.1016/j.cels.2022.12.006.

Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2): 295–307, 2015.

Mengze Du, Yuwei Wang, Li Xie, Gaoliang Deng, Jiansheng Guo, Bo Han, Zhong-Hua Chen, Chen Rui, Jiankang Han, Yuan Chen, Yanru Zhao, Runzhou Cao, Fei Wang, Kun Li, Yu Wang, Yong He, and Xuping Feng. A foundation ai model enhances electron microscopy image analysis. *bioRxiv*, 2026. doi: 10.64898/2026.02.28.708664.

Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI brainlesion workshop*, pp. 272–284. Springer, 2021.

Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 574–584, 2022.

Minkyu Jeon, Rishwanth Raghu, Miro Astore, Geoffrey Woollard, Ryan Feathers, Alkin Kaz, Sonya M Hanson, Pilar Cossio, and Ellen D Zhong. Cryobench: Diverse and challenging datasets for the heterogeneity problem in cryo-em. *Advances in Neural Information Processing Systems*, 37:89468–89512, 2024.

Mohammad Khateri, Morteza Ghahremani, Alejandra Sierra, and Jussi Tohka. Self-supervised super-resolution approach for isotropic reconstruction of 3d electron microscopy images from anisotropic acquisition. In *2023 11th European Workshop on Visual Information Processing (EUVIP)*, pp. 1–6, 2023. doi: 10.1109/EUVIP58404.2023.10323074.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1833–1844, 2021.

Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 136–144, 2017.

Gao Longmi and Gao Pan. Vemamba: Efficient isotropic reconstruction of volume electron microscopy with axial-lateral consistent mamba. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pp. 565–571. Ieee, 2016.

Pawel A Penczek. Reliable cryo-em resolution estimation with modified fourier shell correlation. *IUCrJ*, 7(6):995–1008, 2020.

Stephen M. Plaza and Jan Funke. Analyzing Image Segmentation for Connectomics. *Frontiers in Neural Circuits*, 12:102, November 2018. ISSN 1662-5110. doi: 10.3389/fncir.2018.00102. URL https://www.frontiersin.org/article/10.3389/fncir.2018.00102/full.

Yanjun Qian, Jiaxi Xu, Lawrence F. Drummy, and Yu Ding. Effective Super-Resolution Methods for Paired Electron Microscopic Images. *IEEE Transactions on Image Processing*, 29:7317–7330, 2020. ISSN 1057-7149, 1941-0042. doi: 10.1109/TIP.2020.3000964. URL https://ieeexplore.ieee.org/document/9117049/.

Junchao Qiu, Guojia Wan, Zhengyun Zhou, Minghui Liao, Xiangdong Liu, Xinyuan Li, and Bo Du. Isotropic Reconstruction of Anisotropic vEM Volumes with ViT-Guided Diffusion. *Electronics*, 15(6):1181, March 2026. ISSN 2079-9292. doi: 10.3390/electronics15061181. URL https://www.mdpi.com/2079-9292/15/6/1181.

Erney Ramírez-Aportela, David Maluenda, Yunior C Fonseca, Pablo Conesa, Roberto Marabini, J Bernard Heymann, Jose Maria Carazo, and Carlos Oscar S Sorzano. Fsc-q: a cryoem map-to-atomic model quality validation based on the local fourier shell correlation. *Nature Communications*, 12(1):42, 2021.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, April 2023. doi: 10.1109/TPAMI.2022.3204461.

Louis K Scheffer, C Shan Xu, Michal Januszewski, Zhiyuan Lu, Shin-ya Takemura, Kenneth J Hayworth, Gary B Huang, Kazunori Shinomiya, Jeremy Maitlin-Shepard, Stuart Berg, Jody Clements, Philip M Hubbard, William T Katz, Lowell Umayam, Ting Zhao, David Ackerman, Tim Blakely, John Bogovic, Tom Dolafi, Dagmar Kainmueller, Takashi Kawase, Khaled A Khairy, Laramie Leavitt, Peter H Li, Larry Lindsey, Nicole Neubarth, Donald J Olbris, Hideo Otsuna, Eric T Trautman, Masayoshi Ito, Alexander S Bates, Jens Goldammer, Tanya Wolff, Robert Svirskas, Philipp Schlegel, Erika Neace, Christopher J Knecht, Chelsea X Alvarado, Dennis A Bailey, Samantha Ballinger, Jolanta A Borycz, Brandon S Canino, Natasha Cheatham, Michael Cook, Marisa Dreher, Octave Duclos, Bryon Eubanks, Kelli Fairbanks, Samantha Finley, Nora Forknall, Audrey Francis, Gary Patrick Hopkins, Emily M Joyce, SungJin Kim, Nicole A Kirk, Julie Kovalyak, Shirley A Lauchie, Alanna Lohff, Charli Maldonado, Emily A Manley, Sari McLin, Caroline Mooney, Miatta Ndama, Omotara Ogundeyi, Nneoma Okeoma, Christopher Ordish, Nicholas Padilla, Christopher M Patrick, Tyler Paterson, Elliott E Phillips, Emily M Phillips, Neha Rampally, Caitlin Ribeiro, Madelaine K Robertson, Jon Thomson Rymer, Sean M Ryan, Megan Sammons, Anne K Scott, Ashley L Scott, Aya Shinomiya, Claire Smith, Kelsey Smith, Natalie L Smith, Margaret A Sobeski, Alia Suleiman, Jackie Swift, Satoko Takemura, Iris Talebi, Dorota Tarnogorska, Emily Tenshaw, Temour Tokhi, John J Walsh, Tansy Yang, Jane Anne Horne, Feng Li, Ruchi Parekh, Patricia K Rivlin, Vivek Jayaraman, Marta Costa, Gregory Sxe Jefferis, Kei Ito, Stephan Saalfeld, Reed George, Ian A Meinertzhagen, Gerald M Rubin, Harald F Hess, Viren Jain, and Stephen M Plaza. A connectome and analysis of the adult drosophila central brain. *eLife*, 9:e57443, September 2020. ISSN 2050-084X. doi: 10.7554/eLife.57443. URL https://elifesciences.org/articles/57443.

Amit Suveer, Anindya Gupta, Gustaf Kylberg, and Ida-Maria Sintorn. Super-resolution reconstruction of transmission electron microscopy images using deep learning. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pp. 548–551. IEEE, 2019. ISBN 978-1-5386-3641-1. doi: 10.1109/ISBI.2019.8759153. URL https://ieeexplore.ieee.org/document/8759153/.

Yucheng Tang, Dong Yang, Wenqi Li, Holger R. Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20698–20708, 2022. doi: 10.1109/CVPR52688.2022.02007.

Marin Van Heel and Michael Schatz. Fourier shell correlation threshold criteria. *Journal of structural biology*, 151(3):250–262, 2005.

Eric J Verbeke, Marc Aurèle Gilles, Tamir Bendory, and Amit Singer. Self fourier shell correlation: properties and application to cryo-et. *Communications Biology*, 7(1):101, 2024.

Zhou Wang and Alan C Bovik. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE signal processing magazine*, 26(1):98–117, 2009.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Martin Weigert, Uwe Schmidt, Tobias Boothe, Andreas Müller, Alexandr Dibrov, Akanksha Jain, Benjamin Wilhelm, Deborah Schmidt, Coleman Broaddus, Siân Culley, et al. Content-aware image restoration: pushing the limits of fluorescence microscopy. *Nature methods*, 15(12):1090–1097, 2018.

C. Shan Xu, Song Pang, Gleb Shtengel, Andreas Müller, Alex T. Ritter, Huxley K. Hoffman, Shin-ya Takemura, Zhiyuan Lu, H. Amalia Pasolli, Nirmala Iyer, Jeeyun Chung, Davis Bennett, Aubrey V. Weigel, Melanie Freeman, Schuyler B. Van Engelenburg, Tobias C. Walther, Robert V. Farese, Jennifer Lippincott-Schwartz, Ira Mellman, Michele Solimena, and Harald F. Hess. An open-access volume electron microscopy atlas of whole cells and tissues. *Nature*, 599(7883):147–151, November 2021. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-021-03992-4. URL https://www.nature.com/articles/s41586-021-03992-4.

Jibin Yang, Jie Huo, Muyu Liu, Chenjie Feng, Yan Zhang, Gang Pan, Wenjia Meng, and Renmin Han. veminr: Ultra-fast isotropic reconstruction for volume electron microscopy with implicit

neural representation. *Advanced Science*, 13(13):e11922, 2026. doi: https://doi.org/10.1002/advs.202511922. URL https://advanced.onlinelibrary.wiley.com/doi/abs/10.1002/advs.202511922.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018a.

Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 286–301, 2018b.

# A BENCHMARK CONSTRUCTION

The datasets benchmarked and details are summarized in Table A1.

Table A1: Overview of benchmark datasets. Biological context and technical specifications of the four volumetric datasets used in this benchmark. Cross-resolution consistency was assessed across all compared slice pairs using SSIM and MSE.

| **OpenOrganelle ID & Biological Source** | **Sample Preparation** | **Imaging Parameters** | **Cross-resolution Consistency**[1] | **Key Common Subcellular Structures** |
|---|---|---|---|---|
| `jrc_mus-epididymis`, Mouse Epididymis | Chemical fixation | N/A[2] | SSIM: $0.9970 \pm 0.0007$ MSE: $1.0486 \pm 0.1396$ | Mitochondria, Nucleus, ER, Golgi apparatus, Vesicles |
| `jrc_mus-pancreas`, Mouse Pancreas | High-pressure freezing | LC: 700 eV BC: 0.14 nA SR: 0.2 MHz | SSIM: $0.9674 \pm 0.0044$ MSE: $0.1139 \pm 0.0042$ | |
| `jrc_fly-mb`, Drosophila Mushroom Body | HPF freeze substitution | LC: 700 eV BC: 0.25 nA SR: 0.5 MHz | SSIM: $0.9955 \pm 0.0009$ MSE: $2.0776 \pm 0.4704$ | |
| `jrc_fly-vnc`, Drosophila Ventral Nerve Cord | Chemical fixation PLT-LTS | LC: 900 eV BC: 0.25 nA SR: 0.2 MHz | SSIM: $0.9806 \pm 0.0032$ MSE: $0.8519 \pm 0.0230$ | |

[1] Higher SSIM and lower MSE indicate better.
[2] Imaging parameters were not available for this dataset.

## A.1 COORDINATE MAPPING BETWEEN S1 AND S0

$$x = \alpha_1 p_1 + \beta_1, \qquad p_0 = \frac{x - \beta_0}{\alpha_0}. \tag{A1}$$

where $p_1$ denotes the pixel coordinate in $S_1$, $x$ denotes the coordinate in the shared reference system, and $p_0$ denotes the corresponding pixel coordinate in $S_0$. $\alpha_1$ and $\alpha_0$ represent the scaling parameters for $S_1$ and $S_0$, respectively, while $\beta_1$ and $\beta_0$ represent the corresponding translation parameters.

## A.2 VOXEL CLASS IMBALANCE

vEM volumes are spatially heterogeneous, with a small number of major structures occupying a large fraction of the total voxels. As shown in Figure A1, in the representative dataset jrc_fly-mb, cytosol, Golgi lumen, and Golgi membrane together comprise approximately 80% of the total voxels, whereas mitochondrial membrane accounts for only 2.96%. This distribution motivates the use of informative region selection rather than uniform sampling for benchmark construction.

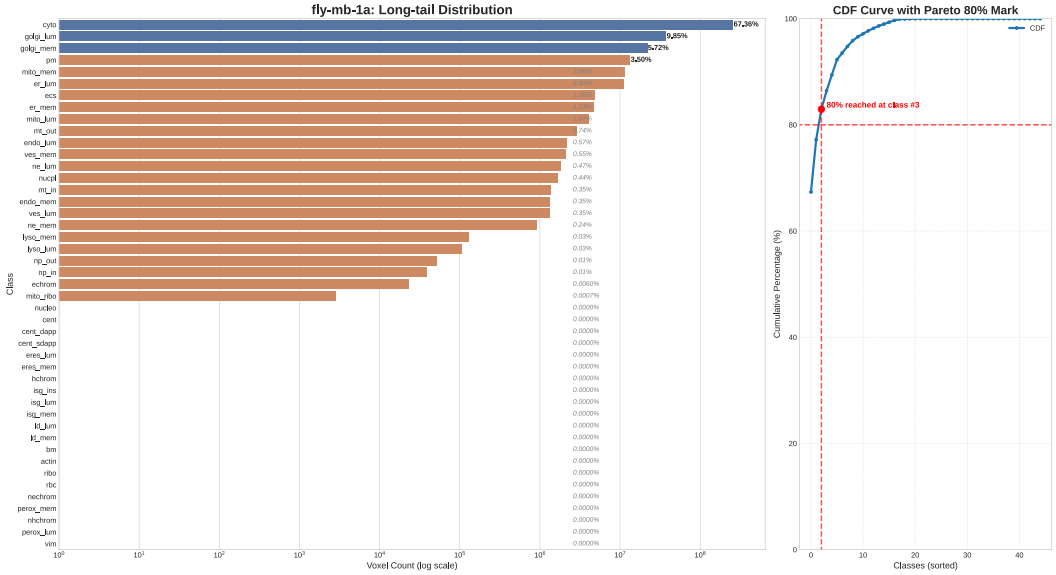


Figure A1: Class distribution in the representative dataset jrc_fly-mb. The cumulative distribution illustrates the strong imbalance among voxel classes, with several major structures accounting for most of the volume.

# B BENCHMARK PROTOCOL

## B.1 DEGRADATION PARAMETERS

The low-resolution benchmark inputs were generated using the degradation pipeline described in the main text. Detailed degradation components and parameter settings are provided in Table A2.

Table A2: Degradation parameters for generating low-resolution benchmark training inputs.

| **Component** | **Description** | **Parameter Settings** |
|---|---|---|
| Blur | Gaussian blur | Kernel size = 5; Blur $\sigma = 0.8$ |
| Downsampling | Resolution reduction | Scale factor = 2 |
| Poisson noise | Shot noise | $k = 200$ |
| Gaussian noise | Additive noise | Noise $\sigma = 0.01$ |
| Reproducibility | Random seed | Seed = 42 |

## B.2 MODEL SOURCES AND ADAPTATION

Detailed implementation choices for each model are summarized in Table A3.

Table A3: Model sources and principal task-specific adaptations for 3D isotropic SR.

| **Model** | **Source Architecture/Formulation** | **Task-specific Adaptation** | **Key Design** |
|---|---|---|---|
| `ResUNet3D` | 3D U-Net (MONAI) Çiçek et al. (2016) | Residual volumetric SR | Normalization-free blocks; zero-initialized output |
| `SwinUNETR3D` | Swin UNETR (MONAI) Tang et al. (2022) | Volumetric SR head | 3D convolutions; pixel shuffle |
| `LIIF3D` | LIIF Chen et al. (2021) | 3D implicit reconstruction | 3D feature unfolding; relative-coordinate querying |
| `Diffusion3D` | Conditional diffusion for SR Saharia et al. (2023) | Packed-domain volumetric SR | $v$-prediction; 3D pixel rearrangement |

## B.3 BENCHMARKING POLICY

The training setup is summarized in Table A4.

Table A4: Unified implementation parameters and benchmarking policy.

| **Configuration Category** | **Benchmarking Specification** |
|---|---|
| Parameter Budget | $\approx$ 6M |
| Data Partition | Predefined split (70% Train / 15% Val / 15% Test) |
| Spatial Input Tile | $64 \times 64 \times 64$ voxels (Isotropic) |
| Effective Batch Size | 8 (Physical BS = 2, Accumulation Steps = 4) |
| Optimization | Adam / AdamW, Gradient Clipping = 1.0 |
| Learning Rate Schedule | Cosine Annealing / MultiStepLR (Model-native) |
| Training Budget | 200K iterations |
| Hardware | Single NVIDIA A100 GPU (40 GB) |

## B.4 CALCULATION OF VOLUME PRESERVATION

Volume preservation was quantified by the absolute relative volume error $|E_{\mathrm{rel}}|$:

$$E_{\mathrm{rel}} = \frac{N_{\mathrm{SR}} - N_{\mathrm{HR}}}{N_{\mathrm{HR}}}, \tag{A2}$$

where $N_{\mathrm{SR}}$ and $N_{\mathrm{HR}}$ denote the foreground voxel counts in the SR and HR masks, respectively. Empty SR masks paired with non-empty HR references were assigned $|E_{\mathrm{rel}}| = 1.0$ and treated as downstream failures. Statistical significance across methods was assessed using paired Wilcoxon signed-rank tests with Holm correction.

# C IMAGE-LEVEL FIDELITY EVALUATION

## C.1 OVERALL QUANTITATIVE COMPARISON ACROSS FOUR DATASETS

Additional quantitative comparisons among the evaluated models are provided in Table A5.

Table A5: Quantitative comparison on epididymis, pancreas, fly-mb, and fly-vnc datasets. The best results are shown in bold and the second-best results are underlined. Values are presented as Mean (SD).

| | | Reconstruction Quality ↑ | | | Perceptual Quality ↓ |
|---|---|---|---|---|---|
| **Dataset** | **Method** | **PSNR** | **SSIM** | **FSC-AUC** | **LPIPS** |
| **epididymis** | Trilinear | 29.71 (0.39) | 0.7090 (0.0297) | 0.6084 (0.1101) | 0.1432 (0.0304) |
| | ResUNet3DLite | 31.43 (0.81) | 0.7911 (0.0249) | 0.6903 (0.1021) | 0.3427 (0.0588) |
| | ResUNet3D | 32.01 (0.82) | 0.8161 (0.0225) | 0.7252 (0.0897) | 0.3219 (0.0600) |
| | SwinUNETR3D | 31.92 (0.83) | 0.8128 (0.0228) | 0.7229 (0.0898) | 0.3217 (0.0598) |
| | LIIF3D | **32.09 (0.81)** | **0.8187 (0.0228)** | **0.7295 (0.0899)** | 0.3109 (0.0647) |
| | Diffusion3D | 30.14 (0.88) | 0.7502 (0.0296) | 0.6271 (0.1137) | **0.1064 (0.0198)** |
| **pancreas** | Trilinear | 28.03 (0.40) | 0.5571 (0.0418) | 0.4129 (0.1062) | 0.2207 (0.0396) |
| | ResUNet3DLite | 30.57 (5.49) | 0.6324 (0.0823) | 0.4768 (0.1236) | 0.5414 (0.1362) |
| | ResUNet3D | 30.92 (4.75) | 0.6600 (0.0778) | 0.5271 (0.1282) | 0.5263 (0.1361) |
| | SwinUNETR3D | 30.88 (5.22) | 0.6562 (0.0777) | 0.5213 (0.1258) | 0.5166 (0.1340) |
| | LIIF3D | **30.99 (4.26)** | **0.6671 (0.0770)** | **0.5401 (0.1375)** | 0.5179 (0.1397) |
| | Diffusion3D | 29.13 (3.53) | 0.5874 (0.0927) | 0.4068 (0.1206) | **0.1223 (0.0276)** |
| **fly-mb** | Trilinear | 25.94 (0.26) | 0.5912 (0.0415) | 0.5339 (0.0459) | 0.1976 (0.0177) |
| | ResUNet3DLite | 26.61 (0.38) | 0.6199 (0.0420) | 0.5828 (0.0422) | 0.5056 (0.0415) |
| | ResUNet3D | 26.82 (0.40) | 0.6330 (0.0408) | 0.6035 (0.0390) | 0.4949 (0.0412) |
| | SwinUNETR3D | 26.80 (0.40) | 0.6322 (0.0409) | 0.6027 (0.0392) | 0.4833 (0.0413) |
| | LIIF3D | **26.88 (0.39)** | **0.6357 (0.0410)** | **0.6095 (0.0391)** | 0.4930 (0.0485) |
| | Diffusion3D | 25.07 (0.45) | 0.5453 (0.0419) | 0.4678 (0.0402) | **0.1112 (0.0181)** |
| **fly-vnc** | Trilinear | 25.41 (0.05) | 0.3779 (0.0176) | 0.3716 (0.0232) | 0.3601 (0.0247) |
| | ResUNet3DLite | 26.35 (0.15) | 0.3989 (0.0193) | 0.4333 (0.0304) | 0.8198 (0.0255) |
| | ResUNet3D | 26.48 (0.15) | 0.4084 (0.0204) | 0.4641 (0.0281) | 0.8142 (0.0268) |
| | SwinUNETR3D | 26.48 (0.15) | 0.4087 (0.0205) | 0.4631 (0.0283) | 0.8118 (0.0303) |
| | LIIF3D | **26.50 (0.15)** | **0.4099 (0.0209)** | **0.4684 (0.0282)** | 0.8417 (0.0287) |
| | Diffusion3D | 24.70 (0.16) | 0.3176 (0.0191) | 0.3337 (0.0252) | **0.1026 (0.0093)** |

### C.2 Quantitative performance of regression-based 3D SR models across datasets

Additional quantitative comparisons among the evaluated regression models are provided in Table A6.

Table A6: Quantitative performance of regression-based 3D SR models across datasets. Model values are mean (SD). Max-min gap is the range of model means. Kendall's W and FDR q summarize the Friedman omnibus test.

| Dataset | Metric | ResUnet3D | SwinUNETR3D | LIIF3D | Max-min gap | Kendall's W | FDR q |
|---|---|---|---|---|---|---|---|
| epididymis | PSNR (↑) | 32.0063 (0.8216) | 31.9190 (0.8338) | **32.0937 (0.8067)** | 0.1747 | 0.963 | <0.001 |
| | SSIM (↑) | 0.8161 (0.0225) | 0.8128 (0.0228) | **0.8187 (0.0228)** | 0.0059 | 0.875 | <0.001 |
| | FSC-AUC (↑) | 0.7252 (0.0897) | 0.7229 (0.0898) | **0.7295 (0.0899)** | 0.0067 | 0.948 | <0.001 |
| pancreas | PSNR (↑) | 30.9204 (4.7520) | 30.8849 (5.2157) | **30.9908 (4.2563)** | 0.1059 | 0.836 | <0.001 |
| | SSIM (↑) | 0.6600 (0.0778) | 0.6562 (0.0777) | **0.6671 (0.0770)** | 0.0109 | 0.803 | <0.001 |
| | FSC-AUC (↑) | 0.5271 (0.1282) | 0.5213 (0.1258) | **0.5401 (0.1375)** | 0.0188 | 0.948 | <0.001 |
| fly-mb | PSNR (↑) | 26.8193 (0.3975) | 26.8025 (0.3984) | **26.8794 (0.3867)** | 0.0769 | 0.957 | <0.001 |
| | SSIM (↑) | 0.6330 (0.0408) | 0.6322 (0.0409) | **0.6357 (0.0410)** | 0.0035 | 0.764 | <0.001 |
| | FSC-AUC (↑) | 0.6035 (0.0390) | 0.6027 (0.0392) | **0.6095 (0.0391)** | 0.0069 | 0.871 | <0.001 |
| fly-vnc | PSNR (↑) | 26.4843 (0.1517) | 26.4755 (0.1539) | **26.5043 (0.1490)** | 0.0288 | 0.944 | <0.001 |
| | SSIM (↑) | 0.4084 (0.0204) | 0.4087 (0.0205) | **0.4099 (0.0209)** | 0.0016 | 0.429 | <0.001 |
| | FSC-AUC (↑) | 0.4641 (0.0281) | 0.4631 (0.0283) | **0.4684 (0.0282)** | 0.0052 | 0.872 | <0.001 |

# D VOLUMETRIC RECONSTRUCTION ANALYSIS

## D.1 CONTROLLED COMPARISON BETWEEN RESUNET3D AND RESUNET3DLITE

Both ResUNet3D and ResUNet3DLite used the same trilinearly upsampled input and residual-learning framework. They also used the same UNet depth, channel configuration, and input-output resolution settings when configured with the same number of base filters. Their residual branches differed in spatial dimensionality: ResUNet3D processed the complete interpolated volume using 3D convolutions, whereas the Lite variant processed each interpolated XY slice independently using 2D convolutions. Consequently, only ResUNet3D learned residual features jointly across adjacent slices. The models were not parameter-matched because 3D and 2D convolutions have different parameter and computational requirements.

Table A7: Paired comparison between the ResUNet3DLite and ResUNet3D. Values are reported as mean (SD). Mean $\Delta$ is defined as ResUNet3D minus ResUNet3DLite, with bootstrap 95% confidence intervals. The $q$-values were obtained using Wilcoxon signed-rank tests with Benjamini–Hochberg FDR correction.

| Dataset | Metric | `Lite`[a] mean (SD) | `3D`[b] mean (SD) | Mean $\Delta$ [95% CI] | `3D` better n/N (%) | BH-FDR $q$ |
|---|---|---|---|---|---|---|
| epididymis | PSNR (↑) | 31.43 (0.81) | 32.01 (0.82) | 0.58 [0.56, 0.60] | 54/54 (100.0%) | <0.001 |
| | SSIM (↑) | 0.791 (0.025) | 0.816 (0.022) | 0.025 [0.024, 0.026] | 54/54 (100.0%) | <0.001 |
| | FSC-AUC (↑) | 0.690 (0.102) | 0.725 (0.090) | 0.035 [0.031, 0.038] | 54/54 (100.0%) | <0.001 |
| pancreas | PSNR (↑) | 30.57 (5.49) | 30.92 (4.75) | 0.35 [0.17, 0.50] | 73/76 (96.1%) | <0.001 |
| | SSIM (↑) | 0.632 (0.082) | 0.660 (0.078) | 0.028 [0.025, 0.030] | 73/76 (96.1%) | <0.001 |
| | FSC-AUC (↑) | 0.477 (0.124) | 0.527 (0.128) | 0.050 [0.047, 0.053] | 73/76 (96.1%) | <0.001 |
| fly-mb | PSNR (↑) | 26.61 (0.38) | 26.82 (0.40) | 0.21 [0.20, 0.22] | 66/66 (100.0%) | <0.001 |
| | SSIM (↑) | 0.620 (0.042) | 0.633 (0.041) | 0.013 [0.012, 0.014] | 66/66 (100.0%) | <0.001 |
| | FSC-AUC (↑) | 0.583 (0.042) | 0.604 (0.039) | 0.021 [0.020, 0.022] | 66/66 (100.0%) | <0.001 |
| fly-vnc | PSNR (↑) | 26.35 (0.15) | 26.48 (0.15) | 0.13 [0.13, 0.14] | 70/70 (100.0%) | <0.001 |
| | SSIM (↑) | 0.399 (0.019) | 0.408 (0.020) | 0.009 [0.009, 0.010] | 70/70 (100.0%) | <0.001 |
| | FSC-AUC (↑) | 0.433 (0.030) | 0.464 (0.028) | 0.031 [0.030, 0.032] | 70/70 (100.0%) | <0.001 |

*Note.* [a]`Lite` denotes ResUNet3DLite; [b]`3D` denotes ResUNet3D. "`3D` better" reports the number and percentage of paired samples with positive differences (ResUNet3D minus ResUNet3DLite).

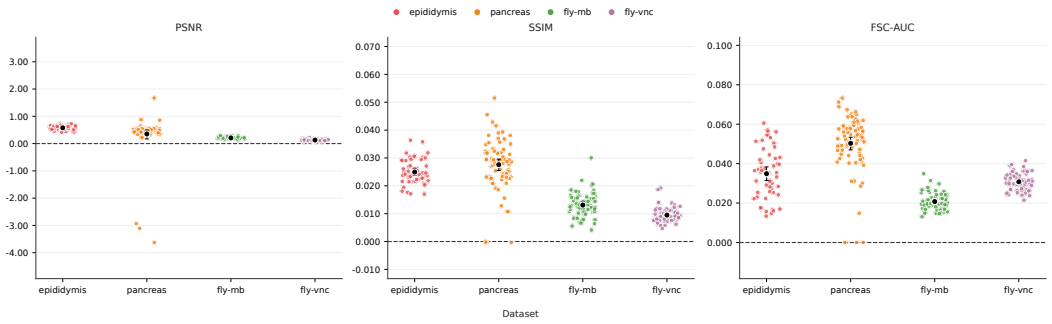


Figure A2: Each dot represents one held-out test-patch difference, computed as ResUNet3D minus ResUNet3DLite; positive values favor ResUNet3D.

### D.2 ADDITIONAL MATERIALS FOR STRUCTURED ARTIFACTS OBSERVATION

Similar patterns were observed across the coronal and sagittal views.

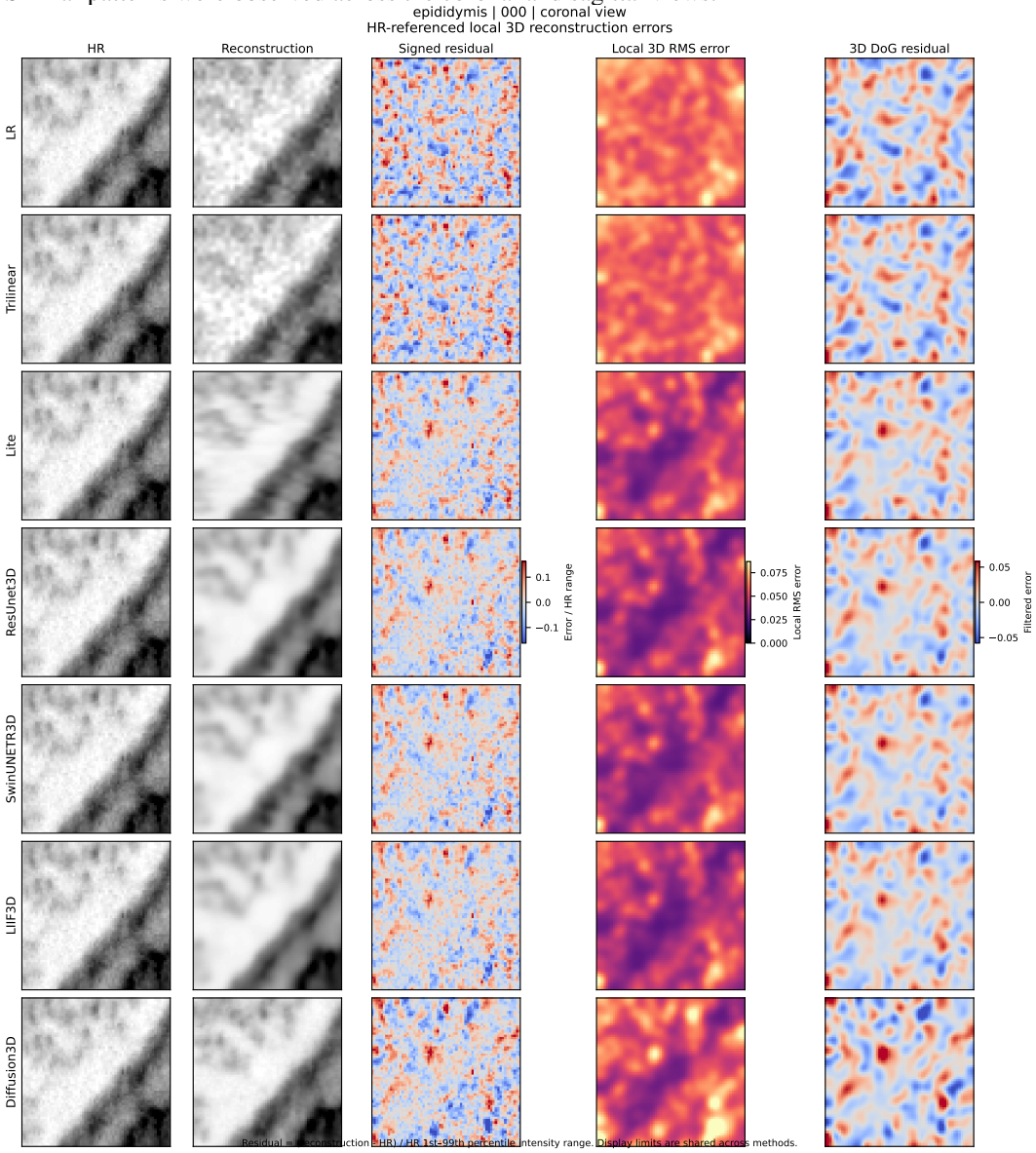


Figure A3: Structured artifacts observed in the reconstructed volumes of the epididymis dataset in the coronal view.

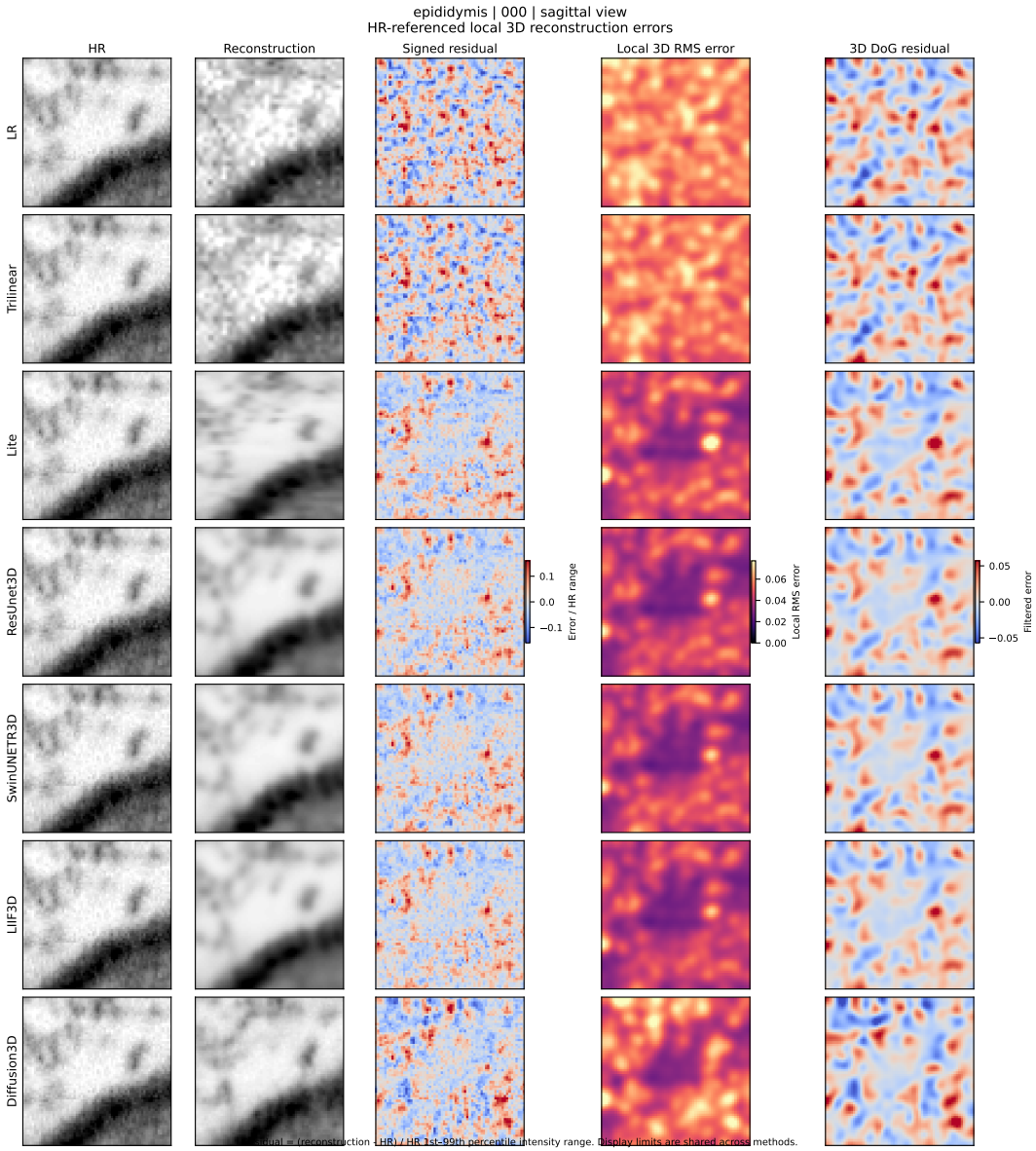


Figure A4: Structured artifacts observed in the reconstructed volumes of the epididymis dataset in the sagittal view, shown for the same sample as above.

# E BIOLOGICAL EVALUATION

Across the 73 paired samples, Diffusion3D had the lowest median absolute relative volume error. Its median absolute foreground voxel-count error was 3,784 voxels (Table A8). The negative association between Dice and absolute relative foreground-volume error was also observed within each reconstruction architecture: Spearman $\rho = -0.570$ for LIIF3D, $\rho = -0.551$ for ResUNet3D, $\rho = -0.538$ for SwinUNETR3D, and $\rho = -0.651$ for Diffusion3D. The association varied across datasets: epididymis ($\rho = -0.766$, $p = 9.6 \times 10^{-13}$), fly-mb ($\rho = -0.749$, $p = 6.0 \times 10^{-12}$), pancreas ($\rho = -0.599$, $p = 3.0 \times 10^{-12}$), and fly-vnc ($\rho = -0.226$, $p = 0.083$). Dice and IoU distributions across the four datasets are shown in Figures A5 and A6.

Table A8: Overall HR-referenced consistency of the segmentation-derived mitochondrial foreground-volume proxy. Each method was evaluated on the same 73 samples. Lower error and lower mean rank indicate better agreement with the HR-derived reference.

| **Method** | **Mean absolute relative error (↓)** | **Median absolute relative error (↓)** | **Mean rank (↓)** | **Best or tied (%, ↑)** |
|---|---|---|---|---|
| `Diffusion3D` | 0.730 | 0.502 | 2.02 | 61.6 |
| `LIIF3D` | 1.658 | 0.938 | 2.51 | 28.8 |
| `ResUNet3D` | 1.629 | 0.799 | 2.65 | 19.2 |
| `SwinUNETR3D` | 1.605 | 1.000 | 2.82 | 23.3 |

### E.1 MEAN DICE AND IOU SCORES

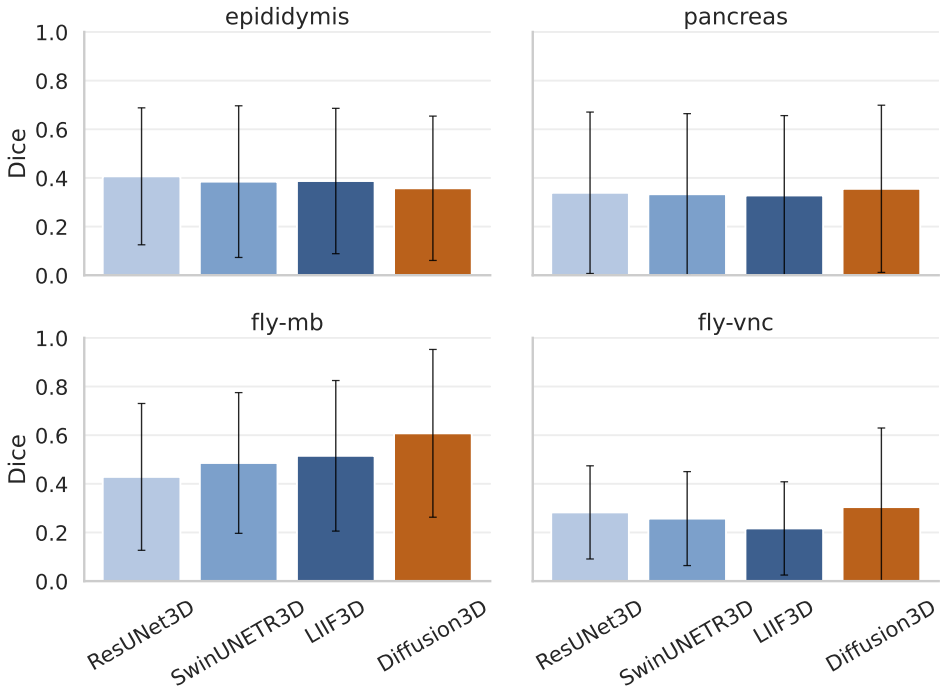


Figure A5: Mean Dice scores of the four evaluated 3D methods across the benchmarked datasets (epididymis, pancreas, fly-mb, and fly-vnc). Higher values indicate greater segmentation overlap consistency.

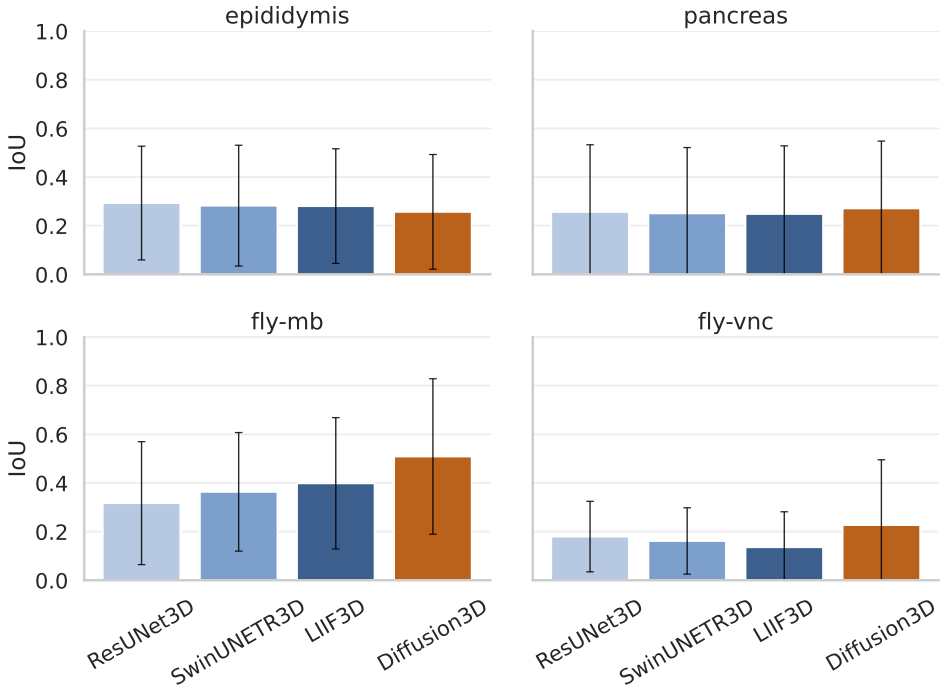


Figure A6: Mean IoU scores of the four evaluated 3D methods across the benchmarked datasets (epididymis, pancreas, fly-mb, and fly-vnc). Higher values indicate greater segmentation overlap consistency.

## F COMPUTATIONAL EFFICIENCY

Table A9: Computational characteristics of the evaluated true 3D SR models.

| Model | Parameters (M) | Peak GPU memory (MB) | Inference time (s) |
|---|---|---|---|
| `ResUNet3D` | 6.00 | 213.4 | 0.01 |
| `SwinUNETR3D` | 5.29 | 500.4 | 0.03 |
| `LIIF3D` | 6.07 | 3664.7 | 3.22 |
| `Diffusion3D` | 5.86 | 302.5 | 1.77 |

*Note.* Runtime and peak GPU memory were measured during inference with a batch size of one and a standardized input volume of $64 \times 64 \times 64$ voxels. All models were evaluated on the same GPU using the same numerical precision. Reported inference time includes the complete reconstruction process. For Diffusion3D, measurements include the full DDIM sampling procedure (50 sampling steps, $\eta = 0.0$). Peak GPU memory denotes the maximum memory allocated by PyTorch during inference.

## G LIMITATION

Our benchmark is tailored to a $2\times$ isotropic reconstruction regime under controlled degradation; the relative efficacy of these methods may shift under different voxel resolutions, noise profiles, or tissue contexts. Downstream validation relies on HR-referenced mitochondrial segmentation; our findings provide evidence of consistency within HR-based analytical pipelines. Extending the evaluation to additional reconstruction regimes, EM-specific structural metrics, and diverse biological tasks remains an avenue for future work.