Spatial-SAM: Spatially Consistent 3D Electron Microscopy Segmentation with SDF Memory and Semi-Supervised Learning

Yikai Huang[1] Renmin Han[2,*] Yuxuan Wang[1] Youcheng Cai[1] Ligang Liu[1,*]

[1]University of Science and Technology of China [2]Shandong University

earendil@mail.ustc.edu.cn hanrenmin@sdu.edu.cn wang42@mail.ustc.edu.cn caiyoucheng@ustc.edu.cn lgliu@ustc.edu.cn

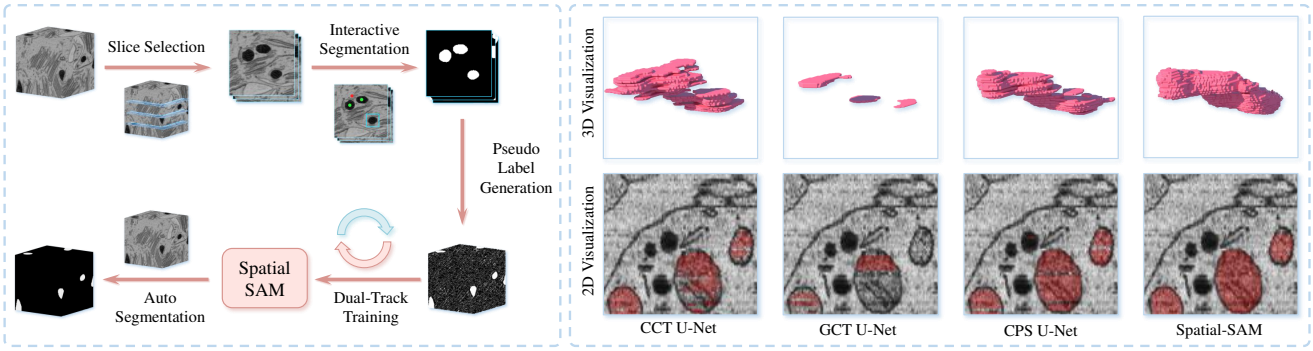

Figure 1. Left: The complete workflow of Spatial-SAM, transitioning from interactive annotation to fully automatic segmentation. Right: Visual comparison of Spatial-SAM and other semi-supervised methods on mitochondria segmentation.

Abstract

Segment Anything Model (SAM)-based approaches have shown strong potential for biomedical image segmentation. However, these methods often struggle to preserve spatial consistency in 3D electron microscopy (3D-EM) data and still require extensive manual annotation. We propose Spatial-SAM, a spatially consistent and annotation-efficient framework for high-precision 3D-EM segmentation. It introduces a 3D Signed Distance Field (SDF) memory mechanism that replaces SAM2's memory with SDF representations precomputed by a 3D U-Net, providing richer geometric information and improving spatial consistency. It also combines SAM2's few-shot capability with a dual-track pseudo-label iterative optimization strategy to learn large-scale 3D-EM segmentation from minimal annotations. Experiments show Spatial-SAM significantly outperforms existing semi-supervised methods and performs comparably to state-of-the-art fully supervised approaches on multiple 3D-EM benchmarks, reducing annotation costs while preserving spatial consistency. Code is available at https://github.com/Giluir/Spatial-SAM.

*Corresponding authors.

1. Introduction

Electron microscopy (EM) provides nanometer-scale views of cellular and tissue ultrastructure, enabling quantitative analysis of membranes, organelles, and synaptic circuitry in neuroscience and cell biology. Modern volume EM modalities, such as serial block-face scanning electron microscopy and focused ion beam scanning electron microscopy [26], routinely produce large-scale and high-resolution 3D EM images, supporting tasks from connectome reconstruction to cellular and organelle morphometry[7, 13]. Across these applications, delineating object boundaries is the gateway to biology: accurate segmentation converts raw contrast into measurable structure, allowing downstream analysis of size, topology, and interaction patterns at scale. In this paper we target this conversion step and focus on automated segmentation that is faithful to 3D morphology while remaining practical for large volumes.

Supervised deep learning has driven progress in EM segmentation. U-Net and its variants remain strong baselines for cells and organelles [2, 27, 29]. To capture longer-range context, Transformer-based designs have been introduced [18, 25, 35], improving global awareness in biomedical images. While these methods perform well on many segmentation tasks, they rely heavily on large amounts of

annotated data, especially when applied to large scale EM datasets. Therefore, semi-supervised and self-supervised learning are proposed to reduce annotation demands. Semi-supervised learning leverages pseudo-labeling with consistency regularization [31], while self-supervised pre-training on large EM corpora improves transfer and data efficiency [6]. Despite these advances, on large 3D EM volumes these methods often falter with sparse labels, appearing reasonable per slice but failing to preserve 3D morphology and leading to inter-slice inconsistencies.

Recently, foundation models for segmentation offer a promising alternative. The general segmentation model Segment Anything (SAM) [15] has demonstrated impressive zero-shot and few-shot capabilities and has been initially applied to biomedical image segmentation [1, 20, 30]. The successor of SAM, SAM2 [28], unifies promptable segmentation for images and videos and incorporates a streaming memory into the model, providing a certain degree of temporal consistency. However, deploying SAM2 on large-scale, high-resolution 3D EM volumes poses two key challenges. On the one hand, maintaining strict spatial continuity and structural integrity across slices of EM volumes is difficult when memory is built from past 2D predictions without explicit volumetric information. On the other hand, achieving robust generalization with minimal manual annotations under significant appearance variation across specimens in large-scale EM datasets remains a challenge.

We address these challenges with Spatial-SAM, a framework that augments SAM2 with geometry-aware volumetric structural guidance and a data-efficient training recipe for large-scale EM. First, we introduce an SDF memory mechanism, where a lightweight 3D U-Net predicts a signed distance field (SDF) over the volume, whose slices provide precomputed structural guidance that condition SAM2 during inference. Compared with logit memories derived from previous predictions, SDF memory encodes object geometry directly, promotes cross-slice smoothness, and avoids error accumulation because it is computed once rather than written from online outputs. Second, we develop a dual-track semi-supervised training scheme that leverages the few-shot capabilities of SAM2 to bootstrap high-quality pseudo-labels and alternates between SDF regression and mask learning to steadily improve both modules. Together, these components form a practical pipeline that starts from a small set of corrected 2D annotations and scales to fully automatic segmentation of large 3D EM volumes. Experiments on several authoritative mitochondria and nuclei datasets indicate that Spatial-SAM achieves competitive or better accuracy than strong semi-supervised and even fully supervised baselines with only 1/64 slices annotated, while delivering markedly improved spatial consistency. Our main contributions are summarized as follows:

- A spatially consistent framework for large-scale 3D EM segmentation that combines the few-shot strengths of SAM2 with volumetric structural guidance.
- An SDF-based memory that replaces the original logit memory with 3D structural cues, providing direction-agnostic, smooth, and error-resilient guidance to improve cross-slice spatial continuity.
- A dual-track semi-supervised training strategy that alternates between SDF regression and mask learning, enabling small 2D annotation budgets to generalize to full volume segmentation.

2. Related Works

2.1. Electron Microscope Image Segmentation

Traditional EM image segmentation primarily relies on thresholding, eigenvector analysis [8], and hierarchical region merging using the watershed algorithm [16], but these methods struggle with complex structures. In recent years, deep learning has become the mainstream approach. End-to-end CNN-based models such as FCN [17] and U-Net [29] have been widely applied to cell and organelle segmentation [27, 29]. The encoder–decoder design with skip connections of U-Net integrates multi-scale context, greatly improving accuracy. For 3D EM data, 3D convolutional networks such as 3D U-Net [5] and VNet [22] directly learn volumetric features but face higher computational cost and limited resolution. Hybrid 2D–3D networks [9] and flood-filling networks (FFN) [12] have also been proposed. With increasing demand for long-range contextual dependencies, Transformer-based models [18, 25, 35] have achieved strong performance by modeling global context, though at high computational expense. Overall, deep learning has improved segmentation accuracy via automatic feature extraction, yet still demands large-scale labels and computation, and high-resolution 3D segmentation remains challenging.

To reduce annotation cost, semi-supervised methods have been explored. Typical strategies include pseudo-label learning and consistency regularization [4, 14, 19, 23]. Building on these ideas, several studies have adapted semi-supervised methods to EM image segmentation. Takaya et al. [31] proposed "4S," which iteratively expands pseudo-labels by leveraging inter-slice correlation. Wolny et al. [33] introduced embedding consistency and push-pull losses to enhance feature separability. Mai et al. [21] developed a "double reliable" network using pixel aggregation and prototype selection for semi-supervised mitochondrial segmentation. In self-supervised pre-training, Conrad and Narayan [6] employed contrastive learning on large unlabeled EM datasets followed by fine-tuning. Although these methods reduce labeling requirements, most are based on 2D slices and struggle to model the spatial continuity of complex 3D structures, often resulting in inter-slice discontinuities or missed fine details.

2.2. SAM-based Methods in Biomedical Imaging

SAM [15] is a general-purpose segmentation foundation model with a Vision Transformer encoder and a prompt-based mask decoder. It supports flexible prompts (points, boxes, text) and demonstrates strong zero/few-shot capability. In biomedical imaging, Ma et al. [20] fine-tuned SAM to create MedSAM, improving medical image segmentation, while Archit et al. [1] introduced μSAM for optical and EM microscopy, achieving better performance across imaging modalities. However, these 2D approaches treat volumetric data as independent slices, limiting spatial correlation modeling.

The recent SAM2 [28] extends SAM with a streaming memory mechanism and large-scale video data, achieving higher accuracy and speed. Shah et al. [30] further implemented 3D memory with momentum updates and low-rank adaptation (LoRA) fine-tuning. Yet, their reliance on prior 2D slice results can propagate local errors, degrading segmentation consistency. Our approach addresses this issue by embedding a 3D object representation into the memory encoding, providing more coherent spatial context.

3. Method

3.1. Preliminary

Segment Anything 2 [28] is a promptable segmentation model for images and videos that inherits the encoder–decoder design of SAM [15] and adds a streaming memory. Given a slice (image) I_t and prompts p (points, boxes, masks), the image encoder produces features $\mathbf{z}_t = \phi(I_t)$ and the prompt encoder yields tokens $\mathbf{e}_p = E(p)$. The decoder predicts k candidate masks with quality scores

$$\{(m_j, s_j)\}_{j=1}^{k} = g\big(\mathbf{z}_t, \mathbf{e}_p\big), \tag{1}$$

and outputs the top mask or a small set of masks. To maintain cross-slice (temporal) consistency, SAM2 writes a compact memory entry from the current embedding and predicted mask into a memory bank $\mathcal{B}$ and, for the next slice, retrieves context via memory attention to condition the features as $\tilde{\phi}(I_{t+1}) = \mathrm{Attn}\big(\phi(I_{t+1}), \mathcal{B}\big)$. With an empty bank, SAM2 naturally reduces to SAM. This unified view lets us use the same promptable interface for fast interactive annotation and provides a foundation to explicitly strengthen cross-slice consistency during propagation across ordered EM slices.

3.2. Overview

We refactor SAM2 into a spatially coherent 3D EM segmentation tool by introducing an SDF memory that supplies geometry-aware volumetric structural guidance. A lightweight 3D U-Net is trained to precompute a SDF to encode memory. Around this model, we build a workflow that scales sparse annotation to full automation (Fig. 2). From a large EM dataset $\mathcal{D}_{\text{all}}$ we select a subset $\mathcal{D}$ and interactively annotate only m slices with SAM2 to obtain high-quality masks $\{Y_j\}_{j=1}^{m}$. These conditional frames seed SAM2 to propagate and produce initial pseudo-labels $\tilde{Y}$ across $\mathcal{D}$. Training then alternates two tracks in a few-shot–guided semi-supervised scheme—one of our key innovations: We utilize the dual-track consistency between the SDF generation of our 3D U-Net module and the mask prediction of the SAM2 module to iteratively improve the quality of both tracks. The trained model can then be applied to the entire dataset $\mathcal{D}_{\text{all}}$ for fully automatic segmentation. The overall architecture is shown in Fig. 2.

3.3. SDF Memory Mechanism

SDF Memory for Spatial Consistency. Our goal is to extend SAM2 to fully automatic 3D volume segmentation while maintaining spatial consistency in the results. A straightforward strategy is to slice the volume along a chosen axis and propagate frame by frame through the memory mechanism of SAM2. Formally, let the input volume be $V \in \mathbb{R}^{D\times H\times W}$, and slicing along that axis yields $\{I_1, I_2, \ldots, I_D\}$. When segmenting I_t, SAM2 can encode features from $\{I_{t-k}, \ldots, I_{t-1}\}$ together with mask logits as memory. However, such one-way propagation has notable limitations. First, it is direction-dependent: the memory contains only past frames along the propagation direction and cannot exploit future frames, leading to inconsistent quality across different directions. Second, errors can accumulate: if a frame is mispredicted, the erroneous result is written into the memory bank and amplified in subsequent steps. Third, the approach is sensitive to slice selection; the choice of propagation axis and conditional frames can significantly affect the final segmentation.

To address these issues, we propose a 3D memory mechanism based on a precomputed signed distance field (SDF). The core idea is to replace probabilistic mask logits with a more geometrically constrained SDF as the memory representation, thereby providing SAM2 with more robust and complete 3D structural guidance. Formally, for any point $\mathbf{x} \in \mathbb{R}^3$, the SDF is defined as

$$\mathrm{SDF}(\mathbf{x}) = \begin{cases} +\min_{\mathbf{y}\in\partial\Omega_{\text{obj}}} \|\mathbf{x}-\mathbf{y}\|, & \mathbf{x} \in \Omega_{\text{obj}}, \\ -\min_{\mathbf{y}\in\partial\Omega_{\text{obj}}} \|\mathbf{x}-\mathbf{y}\|, & \mathbf{x} \notin \Omega_{\text{obj}}, \end{cases} \tag{2}$$

where Ω_{obj} denotes the target object volume and $\partial\Omega_{\text{obj}}$ is its boundary; we set positive values inside and negative values outside to align with the logit distribution.

Compared with a logit representation, SDF memory offers two key advantages. First, as an implicit representation of a 3D object, the SDF provides a more complete semantic description of the segmentation target in 3D, enabling stronger global perception during slice segmentation.

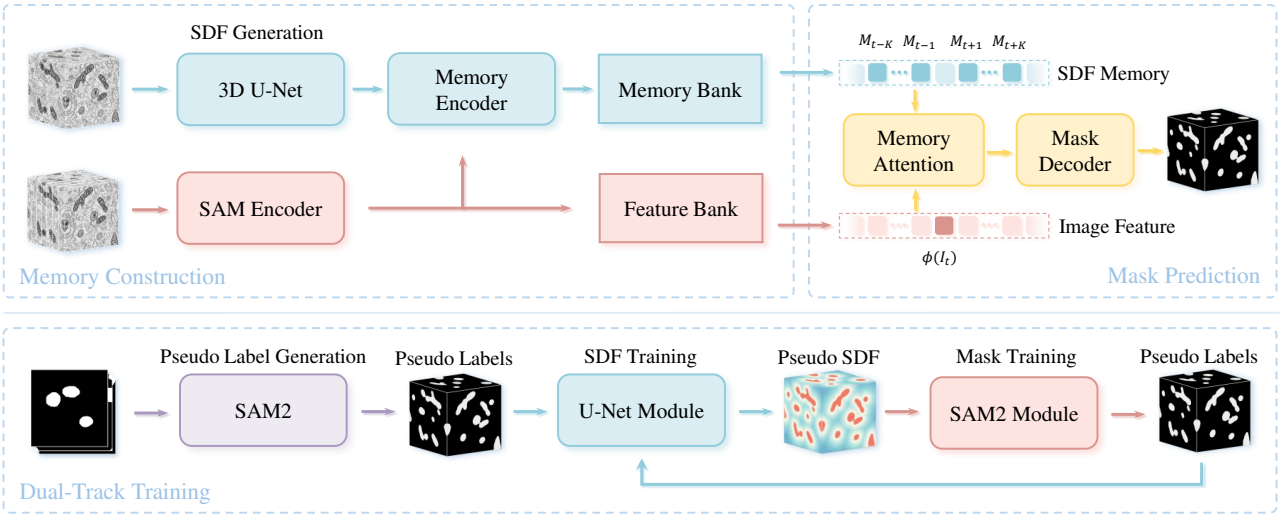

Figure 2. Overview of the proposed Spatial-SAM framework. The upper panel presents the Spatial-SAM model, which extends SAM2 by integrating the SDF Memory mechanism for enhanced spatial representation. The lower panel depicts the proposed dual-track semi-supervised training scheme, which alternates between SDF training and mask training.

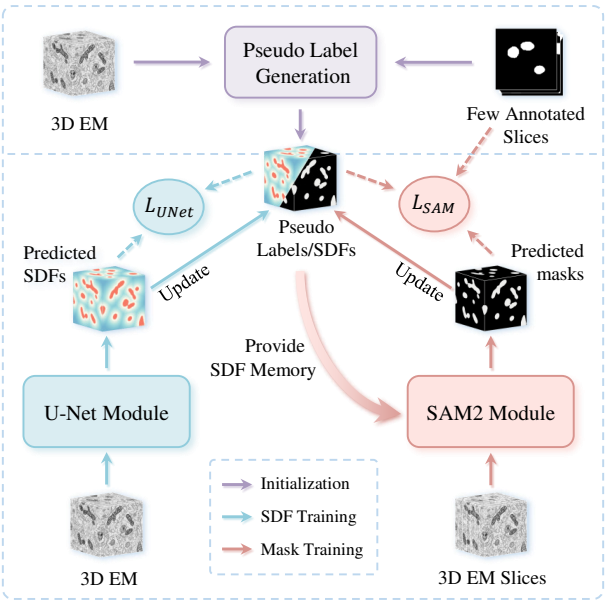

Figure 3. The dual-track training process of Spatial-SAM. The upper part illustrates the initialization process for generating pseudo labels, while the lower part shows the iterative training process, with SDF training on the left and mask training on the right.

Second, the distance values contained in the SDF are naturally spatially smooth. Even with a limited memory length, the model can capture geometric consistency across slices by using the SDF of adjacent frames, markedly reducing spatial discontinuities. Because the SDF is precomputed prior to segmentation, errors from individual slices are not written into memory, thereby avoiding error accumulation. Moreover, SDF memory can serve as a preset contextual cue, allowing SAM2 to complete automatic segmentation without manual prompting during inference.

Learning 3D SDF Memory via U-Net. To obtain reliable SDF memory, we design a lightweight 3D U-Net that serves as a coarse volumetric predictor of geometric structures. The network follows a standard 4-layer encoder–decoder architecture with skip connections. Unlike conventional segmentation networks that output voxel-wise class probabilities, our 3D U-Net is trained to regress an SDF grid, thereby embedding richer geometric and spatial guidance into the representation. Specifically, the network takes as input a downsampled version of the 3D volume $V' \in \mathbb{R}^{D' \times H' \times W'}$ and predicts an SDF grid $\hat{S}' \in \mathbb{R}^{D' \times H' \times W'}$ at low resolution, which is then upsampled to the original resolution. The input volume $V \in \mathbb{R}^{D \times H \times W}$ and the upsampled $\hat{S} \in \mathbb{R}^{D \times H \times W}$ are sliced along the same axis. The image slice I_t is passed through the SAM2 image encoder to obtain a feature representation $\phi(I_t)$. The corresponding SDF slice, together with $\phi(I_t)$, is fed into the memory encoder to generate the memory representation of the slice, M_t, which is stored in the memory bank. When segmenting the target slice I_t, we retrieve the K neighboring entries and condition the current features via memory attention:

$$\tilde{\phi}(I_t) = \mathrm{Attn}\Big(\phi(I_t),\, \{M_\tau\}_{\tau \in \mathcal{N}_t}\Big), \quad \mathcal{N}_t = \{t-K, \ldots, t-1,\, t+1, \ldots, t+K\}. \tag{3}$$

To prevent the coarse SDF prediction from directly interfer-

ing with the segmentation of the current slice, we exclude the memory corresponding to I_t itself and utilize only the SDF memories of neighboring slices.

3.4. Dual-Track Semi-Supervised Training

Large-scale 3D EM volumes exhibit substantial appearance variability across specimens and acquisition sessions, making fully supervised training prohibitively expensive. Our semi-supervised recipe is designed around two complementary ideas: (1) quickly obtaining high-quality initial supervision with minimal annotation budget by exploiting the few-shot capability of SAM2; and (2) enforcing a geometry–semantics agreement across the volume via an alternating dual-track consistency between SDF prediction and mask prediction. Conceptually, this design is inspired by Dual-task Consistency (DTC) [19], which leverages consistency between two tasks for semi-supervised learning. Unlike DTC, which explicitly instantiates two parallel branches, our two tracks are intrinsic to Spatial-SAM: a 3D U-Net that regresses an SDF and SAM2 that decodes masks. They are executed sequentially and coupled through the SDF memory and pseudo-label conversions, enabling iterative training while still benefiting from cross-track consistency. The overall initialization and subsequent alternating refinement process are illustrated in Fig. 3.

Few-shot bootstrapping with SAM2. Departing from conventional pseudo-label pipelines that train a model from scratch (or from a pre-trained model on another dataset) to produce labels, we directly harness the few-shot ability of SAM2 to propagate the annotated slices which serve as conditional frames to nearby slices via the memory bank. This produces high-quality initial pseudo-labels $\tilde{Y}$ with very few annotated slices $\{Y_j\}_{j=1}^{m}$, substantially easing optimization and making convergence easier.

Alternating dual-track consistency. To fully capitalize on unlabeled data, we impose consistency between geometry and semantics by alternating optimization of two tracks: a 3D U-Net that learns a volumetric SDF and SAM2 that predicts 2D masks conditioned on SDF memory. The two tracks exchange supervision signals through conversions between masks and SDF and are trained in a loop:

(1) *SDF training.* Convert pseudo-labels $\tilde{Y}$ into a 3D SDF S and supervise the 3D U-Net to regress $\hat{S}$. This step distills semantic cues from masks into a smooth, direction-agnostic geometric field.

(2) *Mask training.* Slice the predicted $\hat{S}$ to form SDF memory and derive refined pseudo-labels $\tilde{Y}'$ for slices lacking ground truth. Train SAM2 module with $\tilde{Y}'$ together with the few annotated slices $\{Y_j\}_{j=1}^{m}$.

(3) *Iterative refinement.* Re-infer SAM2 over the training subset (or the full dataset) to obtain improved pseudo-labels $\tilde{Y}^{(t+1)}$ and repeat. Each cycle strengthens the U-Net module for SDF and the SAM2 module for masks in tandem.

In this way, the U-Net module continuously learns more accurate 3D geometry, while the SAM2 module improves global segmentation performance driven by high-quality pseudo-labels. The two complement each other. It is worth noting that SAM2 uses a memory mechanism to fuse the 3D geometric representation of U-Net with its own high-resolution feature representation. Therefore, during training, we sample annotated slices and their neighbors with probability p and any slice with probability $1-p$, ensuring full utilization of Ground Truth and preventing the accumulation and propagation of pseudo-label errors. Furthermore, when segmenting a slice, SAM2 only utilizes the SDF memory of its K preceding and succeeding slices as described in Sec. 3, excluding the current one, to avoid over-reliance on the SDF information generated by U-Net. The loss functions for the two modules are designed as follows:

Loss functions. The 3D U-Net regresses a 3D SDF with the following objective:

$$\mathcal{L}_{\text{U-Net}} = \mathcal{L}_{\text{MSE}}(\hat{S}, S) + \lambda \mathcal{L}_{\text{Eikonal}}, \tag{4}$$

where $\mathcal{L}_{\text{MSE}}$ denotes the mean squared error (MSE), used to supervise the numerical consistency of the predicted SDF and the true SDF, while $\mathcal{L}_{\text{Eikonal}}$ constrains the prediction field to satisfy $\|\nabla \hat{S}(\mathbf{x})\| \approx 1$, ensuring the geometric rationality of the SDF. Specifically, the latter is defined as:

$$\mathcal{L}_{\text{Eikonal}} = \frac{1}{|\Omega_{\text{dom}}|} \sum_{\mathbf{x} \in \Omega_{\text{dom}}} \Big(\|\nabla \hat{S}(\mathbf{x})\| - 1 \Big)^2 . \tag{5}$$

For SAM2, we retain its original multi-scale mask prediction and geometry-aware features but replace the interactive prompt with SDF memory. The loss $\mathcal{L}_{\text{SAM2}}$ is computed per slice using the prediction $\hat{Y}_t$ and a hybrid target Y_t^*: $Y_t^* = Y_t$ if slice t is annotated, otherwise $Y_t^* = \tilde{Y}_t'$ (refined pseudo-label). The objective is:

$$\mathcal{L}_{\text{SAM2}}(\hat{Y}_t, Y_t^*) = \alpha \mathcal{L}_{\text{Dice}} + \beta \mathcal{L}_{\text{IoU}} + \gamma \mathcal{L}_{\text{Focal}}, \tag{6}$$

where $\mathcal{L}_{\text{Dice}}$ denotes the Dice Similarity Coefficient (Dice) loss and $\mathcal{L}_{\text{IoU}}$ denotes the Intersection-over-Union (IoU) loss, both encouraging overlap between the predicted and true masks, and in practice $\mathcal{L}_{\text{Focal}}$ addresses foreground/background class imbalance by reducing the weight of easily classified samples.

Through this design, the 3D U-Net focuses on learning a stable volumetric geometry, while SAM2 learns precise masks conditioned on that geometry. The two tracks reinforce each other and culminate in high-quality, fully automatic 3D EM segmentation.

4. Experiments

4.1. Implementation Details

Unless noted, 3D volumes are resized to near-isotropic resolution and cropped to 1024^3. For Spatial-SAM, the memory

Table 1. **Comparison of semantic segmentation performance on different datasets (Dice and mIoU, %).** "Labels" denotes the proportion of labeled slices used. Fully-supervised methods (using Full labels) are listed at the top as the performance upper bound. The **best** results among **semi-supervised** methods are highlighted in **bold**.

Method	Labels	OOMLM		OOMLN		MitoEM-R		MitoEM-H	
		Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU
3D U-Net [5]	Full	94.14	88.93	92.91	86.80	92.95	87.09	91.60	84.59
Swin UNETR [10]	Full	96.74	93.68	98.22	96.51	91.51	84.39	81.10	69.92
Cellpose-SAM [24]	Full	96.44	93.13	98.42	96.90	88.91	80.29	85.25	74.61
μSAM [1]	Full	95.93	92.18	93.17	87.48	92.77	86.55	89.09	80.36
SAM4EM [30]	Full	96.75	93.70	96.61	93.46	95.12	90.71	91.10	83.67
CCT U-Net [23]	1/64	91.56	84.80	87.15	77.53	84.52	73.19	66.20	52.74
GCT U-Net [14]	1/64	95.08	90.63	93.65	88.09	88.66	79.67	73.50	58.24
CPS U-Net [4]	1/64	95.74	91.84	69.03	55.91	93.38	87.60	88.77	79.83
Spatial-SAM (Ours)	1/64	**96.51**	**93.25**	**98.14**	**96.34**	**94.45**	**89.51**	**90.10**	**82.02**

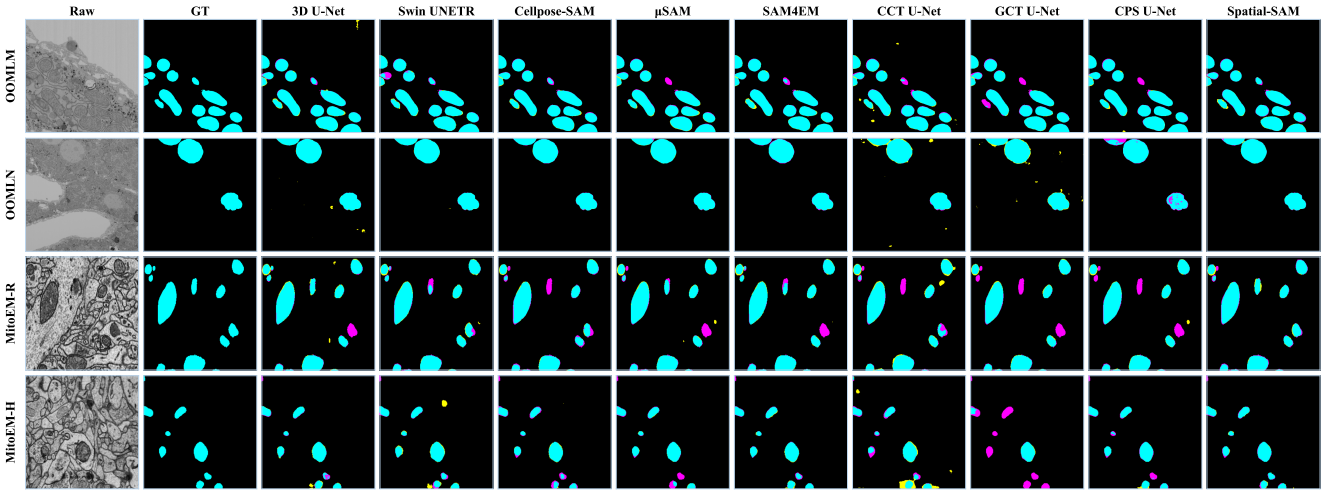

Figure 4. Visualization of segmentation results of the OpenOrganelle and MitoEM datasets. cyan indicates true positives (TP), magenta indicates false negatives (FN), and yellow indicates false positives (FP).

neighborhood size $K = 6$, and the 3D U-Net operates on $1/2$ resolution inputs. For each dataset, we uniformly selected 1/64 2D slices. We used the ground truth annotations of these slices to simulate the results of interactive segmentation using SAM2 along with manual correction, and then used them as conditional frames to generate pseudo-labels for the subsequent semi-supervised training process. Further environmental and hyperparameter details are provided in the supplementary material.

4.2. Datasets and Evaluation Metrics

We conducted extensive experiments on several authoritative 3D EM datasets, covering the segmentation tasks of typical organelles such as mitochondria and nuclei.

OpenOrganelle Dataset. The OpenOrganelle mouse liver dataset [11, 34] provides 3D EM of hepatocytes at an isotropic voxel size of $8\,\mathrm{nm}$ with annotations for mitochondria and nuclei. We work on 1024^3 resolution subvolumes and report results on held-out subvolumes; crop counts and train/validation splits are provided in the supplementary material. For brevity in tables and figures, we refer to *OpenOrganelle mouse liver mitochondria* and *OpenOrganelle mouse liver nuclei* as **OOMLM** and **OOMLN**.

MitoEM Dataset. The MitoEM dataset [32] comprises rat (MitoEM-R) and human (MitoEM-H) EM image stacks, each with 1000 annotated sections (voxel size $30\times8\times8\,\mathrm{nm}$; slice resolution 4096×4096). We train on the official training set and evaluate on the validation set. Further dataset specifics are summarized in the supplementary material.

Evaluation Metrics. We adopt the Dice and the mean Intersection over Union (mIoU) as evaluation metrics for semantic segmentation. All instance segmentation results by

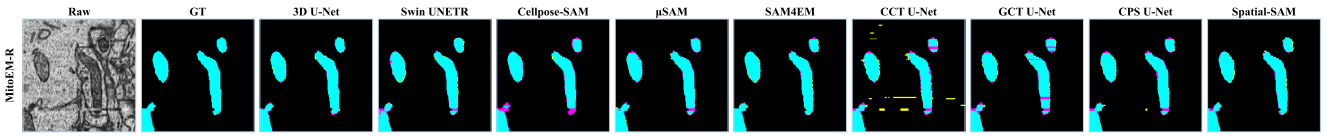

Figure 5. Visualization of segmentation results of the MitoEM-R datasets on x-z plane.

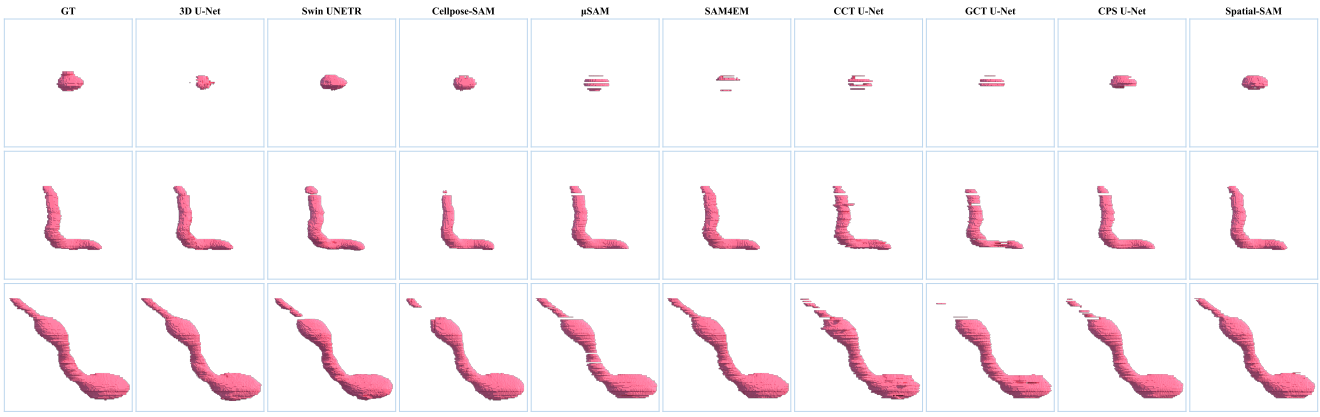

Figure 6. 3D visualization comparison of different methods on mitochondria with varying sizes and morphologies.

some baselines are converted to binary masks and evaluated using the same metrics.

Table 2. Comparison of semi-supervised methods on the MitoEM-R dataset under different partition protocols.

Method	1/64		1/16		1/4	
	Dice	mIoU	Dice	mIoU	Dice	mIoU
CCT U-Net [23]	84.52	73.19	83.79	73.12	87.84	78.47
GCT U-Net [14]	88.66	79.67	91.35	84.10	91.29	84.00
CPS U-Net [4]	93.38	87.60	93.95	88.62	94.19	89.04
Spatial-SAM	**94.45**	**89.51**	**94.91**	**90.33**	**95.35**	**91.12**

Table 3. Ablation study of memory mechanism on MitoEM-R.

Memory Type	Direction	Exclude Self	Dice	mIoU
SAM2	Unidirectional	-	92.62	86.31
SDF	Unidirectional	-	93.62	88.03
SDF	Bidirectional	No	94.11	88.90
SDF	**Bidirectional**	**Yes**	**94.45**	**89.51**

4.3. Comparison with Other Methods

We compare Spatial-SAM with three families of baselines: semi-supervised methods [4, 14, 23] (U-Net backbone, re-trained on the same annotated slices), fully supervised 3D networks [5, 10], and SAM-based approaches [1, 24, 30]. Table 1 reports Dice / mIoU on OOMLM, OOMLN, MitoEM-R, and MitoEM-H.

Under the same 1/64 labeled-slice budget, Spatial-SAM consistently outperforms all semi-supervised baselines across all four datasets. Compared with the strongest

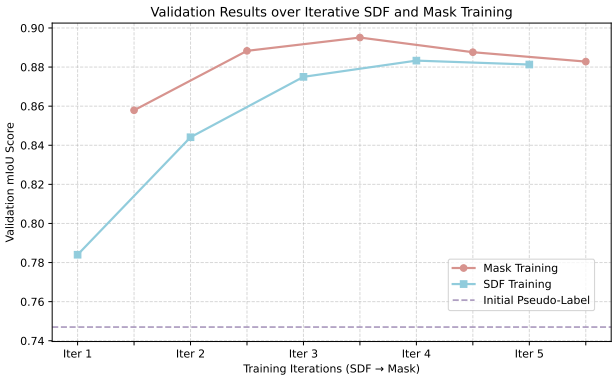

Figure 7. Validation performance across training iterations.

semi-supervised baseline (CPS U-Net), Spatial-SAM improves the four-dataset average by +8.07% Dice and +11.49% mIoU. Such performance stems from the high-quality SAM2-initialized pseudo-labels and the subsequent dual-track alternating refinement.

Compared with the strongest fully supervised baseline (SAM4EM), Spatial-SAM stays at nearly the same four-dataset average level (difference: -0.09% Dice and -0.10% mIoU) while using only 1/64 labels. On large-volume nuclei (OOMLN), 3D U-Net is clearly weaker (92.91% / 86.80%), whereas Spatial-SAM reaches 98.14% / 96.34%; this reflects the advantage of combining SAM2 long-range context with SDF memory for preserving global object structure under sparse supervision. For complex

mitochondria, several methods show evident degradation from MitoEM-R to MitoEM-H, while Spatial-SAM maintains competitive performance with 1/64 labels (90.10% / 82.02%), with smaller performance loss than all semi-supervised baselines.

We further analyze different partition protocols on MitoEM-R (Table 2). Spatial-SAM with only 1/64 labels already surpasses all competing semi-supervised baselines trained with 1/4 labels. As the partition protocol increases from 1/64 to 1/16 and 1/4, Spatial-SAM continues to improve (+0.90% Dice / +1.61% mIoU from 1/64 to 1/4). Taken together, these results show that Spatial-SAM can deliver strong performance even with very sparse slice annotations, and can further benefit from additional labels.

4.4. Visual analysis

Fig. 4 presents qualitative comparisons on OpenOrganelle and MitoEM. Spatial-SAM yields smoother boundaries and reduced noise relative to other semi-supervised methods, with clear benefits in dense, structurally complex regions. Compared with fully supervised baselines, the masks are comparable in fidelity and often cleaner.

Fig. 5 shows the x–z plane on MitoEM-R to assess volumetric coherence. Slice-wise methods frequently display thickness "flicker" and jagged inter-slice transitions; Spatial-SAM maintains consistent cross-slice thickness and suppresses zig-zag artifacts. This behavior is obtained without pre- or post-processing such as z-filtering [3], relying solely on spatially aware propagation. Robustness is especially evident despite acquisition-induced discontinuities or local intensity variations between slices. These observations align with the quantitative gains and illustrate how bidirectional SDF memory suppresses error accumulation and enforces geometry-aware coherence across the volume.

As further illustrated in Fig. 6, 2D semi-supervised and fully supervised models suffer from slice-wise inconsistency, producing discontinuous and fragmented mitochondrial volumes. Pure 3D networks (e.g., 3D U-Net) enhance continuity but can be less adaptable to diverse morphologies. Spatial-SAM integrates the strengths of both paradigms, preserving local 2D precision while maintaining global 3D coherence, and thereby delivering faithful volumetric reconstructions.

4.5. Ablation study

To investigate the effect of the SDF memory mechanism and memory selection, we conduct ablation studies on the **MitoEM-R** dataset. We first compare three variants of memory encoding: baseline original memory, unidirectional SDF memory, and bidirectional SDF memory (Table 3). Introducing SDF-based encoding improves performance, as our approach achieves 93.62% Dice and 88.03% mIoU versus 92.62% / 86.31% for the original representation, demonstrating that continuous signed distance encoding enhances spatial consistency in updates. The bidirectional SDF memory further achieves 94.45% Dice and 89.51% mIoU, surpassing the unidirectional variant by 0.83% / 1.48% and the original representation by 1.83% / 3.20% in Dice/mIoU.

Furthermore, we investigate the choice of neighborhood $\mathcal{N}_t$. As shown by the variant without self-exclusion (the third row in Table 3), including the SDF memory of the current slice I_t ($\{t-K,\ldots,t,\ldots,t+K\}$) actually leads to a performance drop (94.11% Dice). This confirms that excluding the target slice's own coarse SDF prediction prevents self-coupling and error amplification, validating our neighbor-only strategy.

Fig. 7 shows the validation performance across five training iterations. We observe that the model converges quickly within three iterations, with diminishing returns thereafter. Both the mask training and SDF training steadily improve the performance, showing the effectiveness of the dual-task semi-supervised learning framework. Due to the high quality of the initial pseudo-labels generated by SAM2, the model rapidly refines its predictions in the early stages. And in later iterations, the result of mask training and SDF training tend to be closer to each other, indicating that the SDF outputs might have more influence on SAM module during mask training and seems to become a bit overfitted.

4.6. Discussion

Spatial-SAM delivers accurate segmentation and improved spatial consistency without relying on dedicated preprocessing; nevertheless, in extreme cases where acquisition introduces significantly damaged slices or severe exposure inconsistency, the resulting disruption of image slice continuity can still test the limits of inter-slice stability. Moreover, the dual-track semi-supervised procedure—alternating SDF regression and mask learning while regenerating large pseudo-label sets each cycle—incurs higher training time. These limitations arise alongside the gains in label efficiency and spatial consistency, and point to efficiency-oriented refinements as promising future work.

5. Conclusion

Spatial-SAM integrates a SDF representation with a dual-track semi-supervised training strategy built on SAM2, achieving state-of-the-art performance in semantic segmentation of 3D EM volumes. Experiments show that the 3D SDF memory enforces spatial consistency across slices, improving segmentation completeness and accuracy. The combination of few-shot annotation and semi-supervised training yields strong results under extremely sparse supervision. Its robustness as well as scalability are validated on multiple public datasets, making the approach well suited for high-resolution, large-scale biological imaging.

Acknowledgements

We would like to thank the anonymous reviewers for their constructive feedback. This work was supported in part by the National Natural Science Foundation of China (Grant No. 62025207) and by the Dubai Future Foundation (Award No. 2024CANAD-MES-061).

References

[1] Anwai Archit, Luca Freckmann, Sushmita Nair, Nabeel Khalid, Paul Hilt, Vikas Rajashekar, Marei Freitag, Carolin Teuber, Melanie Spitzner, Constanza Tapia Contreras, et al. Segment anything for microscopy. *Nature Methods*, 22(3): 579–591, 2025. 2, 3, 6, 7

[2] Yue Cao, Shigang Liu, Yali Peng, and Jun Li. Denseunet: densely connected unet for electron microscopy image segmentation. *IET Image Processing*, 14(12):2682–2689, 2020. 1

[3] Vincent Casser, Kai Kang, Hanspeter Pfister, and Daniel Haehn. Fast mitochondria detection for connectomics. In *Medical Imaging with Deep Learning*, pages 111–120. PMLR, 2020. 8

[4] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *CVPR*, pages 2613–2622, 2021. 2, 6, 7

[5] 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*, pages 424–432. Springer, 2016. 2, 6, 7

[6] Ryan Conrad and Kedar Narayan. Cem500k, a large-scale heterogeneous unlabeled cellular electron microscopy image dataset for deep learning. *Elife*, 10:e65894, 2021. 2

[7] Julie Faitg, Clay Lacefield, Tracey Davey, Kathryn White, Ross Laws, Stylianos Kosmidis, Amy K Reeve, Eric R Kandel, Amy E Vincent, and Martin Picard. 3d neuronal mitochondrial morphology in axons, dendrites, and somata of the aging mouse hippocampus. *Cell Reports*, 36(6), 2021. 1

[8] Achilleas S Frangakis and Reiner Hegerl. Segmentation of two- and three-dimensional data from electron microscopy using eigenvector analysis. *Journal of Structural Biology*, 138(1):105–113, 2002. 2

[9] Matthias G Haberl, Christopher Churas, Lucas Tindall, Daniela Boassa, Sébastien Phan, Eric A Bushong, Matthew Madany, Raffi Akay, Thomas J Deerinck, Steven T Peltier, et al. Cdeep3m—plug-and-play cloud-based deep learning for image segmentation. *Nature Methods*, 15(9):677–680, 2018. 2

[10] 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 *MICCAI Brainlesion Workshop*, pages 272–284, 2021. 6, 7

[11] Larissa Heinrich, Davis Bennett, David Ackerman, Woohyun Park, John Bogovic, Nils Eckstein, Alyson Petruncio, Jody Clements, Song Pang, C Shan Xu, et al. Whole-cell organelle segmentation in volume electron microscopy. *Nature*, 599(7883):141–146, 2021. 6

[12] Michał Januszewski, Jeremy Maitin-Shepard, Peter Li, Jörgen Kornfeld, Winfried Denk, and Viren Jain. Flood-filling networks. *arXiv preprint arXiv:1611.00421*, 2016. 2

[13] Mandy SJ Kater, Aina Badia-Soteras, Jan RT van Weering, August B Smit, and Mark HG Verheijen. Electron microscopy analysis of astrocyte-synapse interactions shows altered dynamics in an alzheimer's disease mouse model. *Frontiers in Cellular Neuroscience*, 17:1085690, 2023. 1

[14] Zhanghan Ke, Di Qiu, Kaican Li, Qiong Yan, and Rynson WH Lau. Guided collaborative training for pixel-wise semi-supervised learning. In *ECCV*, pages 429–445, 2020. 2, 6, 7

[15] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pages 4015–4026, 2023. 2, 3

[16] Ting Liu, Elizabeth Jurrus, Mojtaba Seyedhosseini, Mark Ellisman, and Tolga Tasdizen. Watershed merge tree classification for electron microscopy image segmentation. In *ICPR*, pages 133–137, 2012. 2

[17] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015. 2

[18] Naisong Luo, Rui Sun, Yuwen Pan, Tianzhu Zhang, and Feng Wu. Electron microscopy images as set of fragments for mitochondrial segmentation. In *AAAI*, pages 3981–3989, 2024. 1, 2

[19] Xiangde Luo, Jieneng Chen, Tao Song, and Guotai Wang. Semi-supervised medical image segmentation through dual-task consistency. In *AAAI*, pages 8801–8809, 2021. 2, 5

[20] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024. 2, 3

[21] Huayu Mai, Rui Sun, Tianzhu Zhang, Zhiwei Xiong, and Feng Wu. Dualrel: Semi-supervised mitochondria segmentation from a prototype perspective. In *CVPR*, pages 19617–19626, 2023. 2

[22] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*, pages 565–571, 2016. 2

[23] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *CVPR*, pages 12674–12684, 2020. 2, 6, 7

[24] Marius Pachitariu, Michael Rariden, and Carsen Stringer. Cellpose-sam: superhuman generalization for cellular segmentation. *bioRxiv*, pages 2025–04, 2025. 6, 7

[25] Yuwen Pan, Naisong Luo, Rui Sun, Meng Meng, Tianzhu Zhang, Zhiwei Xiong, and Yongdong Zhang. Adaptive template transformer for mitochondria segmentation in electron microscopy images. In *ICCV*, pages 21474–21484, 2023. 1, 2

[26] Christopher J Peddie, Christel Genoud, Anna Kreshuk, Kimberly Meechan, Kristina D Micheva, Kedar Narayan, Constantin Pape, Robert G Parton, Nicole L Schieber, Yannick

Schwab, et al. Volume electron microscopy. *Nature Reviews Methods Primers*, 2(1):51, 2022. 1

[27] Tran Minh Quan, David Grant Colburn Hildebrand, and Won-Ki Jeong. Fusionnet: A deep fully residual convolutional neural network for image segmentation in connectomics. *Frontiers in Computer Science*, 3:613981, 2021. 1, 2

[28] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 2, 3

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015. 1, 2

[30] Uzair Shah et al. Sam4em: Efficient memory-based two stage prompt-free segment anything model adapter for complex 3d neuroscience electron microscopy stacks. In *CVPR*. 2, 3, 6, 7

[31] Eichi Takaya, Yusuke Takeichi, Mamiko Ozaki, and Satoshi Kurihara. Sequential semi-supervised segmentation for serial electron microscopy image with small number of labels. *Journal of Neuroscience Methods*, 351:109066, 2021. 2

[32] Donglai Wei, Zudi Lin, Daniel Franco-Barranco, Nils Wendt, Xingyu Liu, Wenjie Yin, Xin Huang, Aarush Gupta, Won-Dong Jang, Xueying Wang, et al. Mitoem dataset: large-scale 3d mitochondria instance segmentation from em images. In *MICCAI*, pages 66–76, 2020. 6

[33] Adrian Wolny, Qin Yu, Constantin Pape, and Anna Kreshuk. Sparse object-level supervision for instance segmentation with pixel embeddings. In *CVPR*, pages 4402–4411, 2022. 2

[34] 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, et al. An open-access volume electron microscopy atlas of whole cells and tissues. *Nature*, 599(7883):147–151, 2021. 6

[35] Xiao Zhang, Zihan Lin, Liguo Wang, Yong S Chu, Yang Yang, Xianghui Xiao, Yuewei Lin, and Qun Liu. Swincell: a 3d transformer and flow-based framework for improved cell segmentation. *Communications Biology*, 8(1):962, 2025. 1, 2

Spatial-SAM: Spatially Consistent 3D Electron Microscopy Segmentation with SDF Memory and Semi-Supervised Learning

Supplementary Material

1. More Implementation Details

Hyperparameters and Environment. We implement Spatial-SAM in PyTorch and run training/inference on a single NVIDIA RTX 3090 GPU. For model optimization, we set the sampling probability of slices with annotations to $p = 0.8$, and the Eikonal loss weight to $\lambda = 0.5$.

Data Augmentation. We also apply a series of data augmentation techniques to improve model robustness. In particular, random brightness adjustment is used to simulate the uneven illumination commonly observed across different regions or slices of EM images, while random flipping, random rotation, and elastic deformation are adopted to enhance the generalization ability of the model to diverse morphological variations.

Toolset Implementation. We provide a full-process toolset based on the Napari plugin, covering:

- **Interactive Segmentation**: Supports fast annotation of 2D or 3D slices using points and box prompts;
- **Model Training**: Trains Spatial-SAM model based on a small number of annotated slices to adapt to new datasets;
- **Fully Automatic Segmentation**: Calls Spatial-SAM to perform semantic/instance segmentation on 3D volumes;
- **Annotation Correction and Retraining**: Supports interactive correction of automatic results and iterative model optimization;
- **Hardware Adaptation**: Allows users to set resolution and memory usage based on device conditions.

Input Resolution. All 3D volumes are processed as subvolumes of size $1024 \times 1024 \times 1024$: high-resolution datasets are partitioned into tiles of this size, while volumes with smaller spatial extents are resampled (and, when necessary, upsampled) to $1024 \times 1024 \times 1024$ in our method.

Few-shot Annotations and Pseudo-Labels. Within each dataset, we uniformly sample 1/64 of the 2D slices that contain foreground objects as few-shot annotations. Ground-truth masks on these slices are used to simulate SAM2-assisted interactive segmentation with light manual correction. The corrected masks are then used as conditional frames to generate pseudo-labels for the remaining slices in the subsequent semi-supervised training.

Baseline Protocols. To ensure fairness, we follow official implementations and training hyperparameters for all baselines. For semi-supervised methods[2, 6, 8], we adopt U-Net as the backbone consistent with their protocols and train using the same few-shot annotated slices as in our method. For baseline 3D methods and SAM-based approaches[1, 3, 4, 9], anisotropic volumes are resampled to approximate isotropy. Other 2D methods are processed slice-by-slice[2, 6, 8]. For all baselines, we adopt the input resolutions recommended by their official implementations.

2. Dataset Details and Splits

OpenOrganelle (Mouse Liver). The OpenOrganelle mouse liver dataset [5, 13] provides complete 3D electron microscopy imaging of hepatocytes, acquired using enhanced focused ion beam scanning electron microscopy (FIB-SEM) with an isotropic voxel resolution of $8\,\mathrm{nm}$. The dataset includes annotations of cellular structures such as mitochondria and nuclei. We constructed the mitochondrial and nucleus segmentation datasets by cropping 14 and 9 subvolumes with a voxel size of $1024 \times 1024 \times 1024$. For the mitochondrial dataset we used the first 9 subvolumes for training and the remaining 5 for validation; for the nucleus dataset we used the first 4 for training and the remaining 5 for validation.

MitoEM. The MitoEM dataset [12] contains two sets of volumetric images, one from rat (MitoEM-R) and one from human (MitoEM-H) tissue. Each volume covers $30 \times 30 \times 30\,\mu m^3$ at a voxel resolution of $30 \times 8 \times 8\,\mathrm{nm}$, comprising 1000 consecutive electron microscopy sections with precise mitochondrial instance annotations. The original training, validation, and test sets follow a $4{:}1{:}5$ split. The spatial resolution of each slice is 4096×4096. Ground-truth annotations are publicly available for the training and validation sets. In our experiments, we use the official training set for training and the validation set for evaluation.

3. Supplementary Results

Fig. S1 and S2 provide supplementary visualizations that complement Fig. 5 and Fig. 6 from the main paper, respectively.

Additional Evaluation Metrics. Table S1 reports voxel-wise precision and recall, complementing the Dice/mIoU results in the main paper. Across all datasets, Spatial-SAM achieves consistently high precision while maintaining strong recall. For instance, on MitoEM-R it reaches 96.40% precision and 92.62% recall, improving recall by +2.68 points over μSAM and by +8.61 points over Cellpose-SAM, indicating fewer missed mitochondria without inflating false positives. Unlike some semi-supervised approaches that can exhibit a pronounced precision–recall imbalance on particular benchmarks (e.g., CPS U-Net attains 99.58% precision but only 56.08% recall on OOMLN),

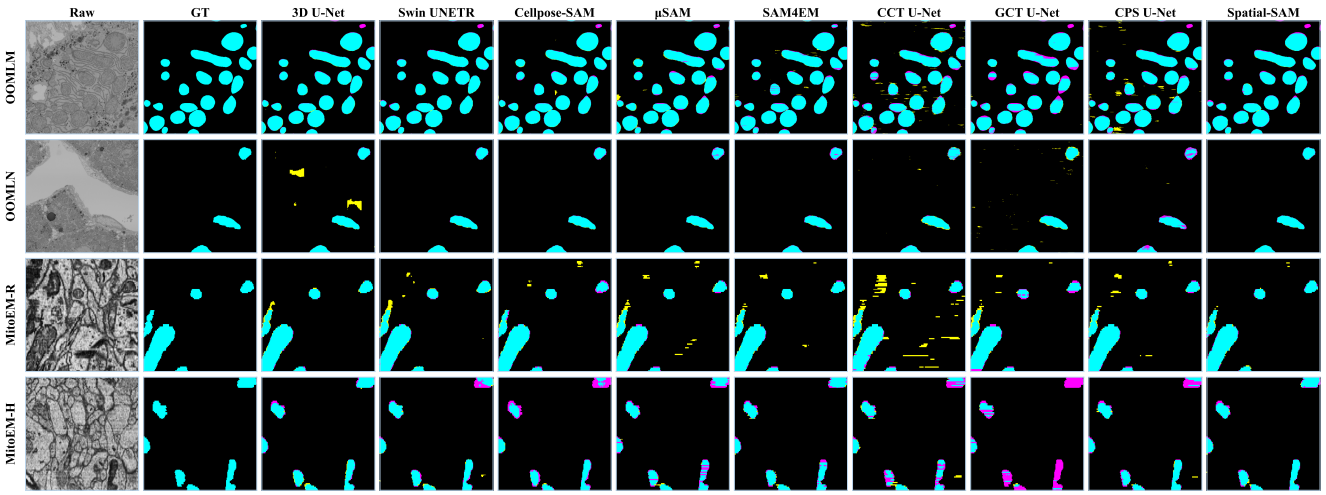

Figure S1. Supplementary visualization of segmentation results on x-z plane.

Figure S2. Supplementary 3D visualization comparison of different methods on mitochondria.

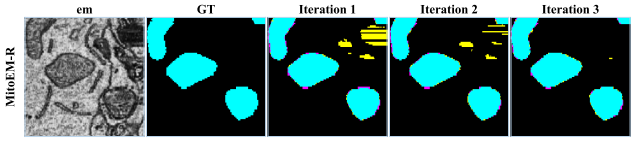

Figure S3. Pseudo-label evolution on MitoEM-R across training iterations. As training proceeds, false positives and false negatives are progressively reduced, indicating improved pseudo-label quality.

Spatial-SAM preserves a more consistent balance between precision and recall across all evaluated datasets, demonstrating its robustness with limited supervision.

Table S2 evaluates boundary quality using the average symmetric surface distance and the 95th-percentile Hausdorff distance (HD95), two standard surface-based metrics for biomedical image segmentation [11]. The average surface distance measures the mean bidirectional distance (in nm) between predicted and reference boundaries, while HD95 summarizes the worst-case deviations (in nm) after discarding the most extreme 5% of surface outliers, which

Table S1. **Comparison of precision and recall on different datasets (%).** The **best** and second-best results for each metric are highlighted.

Method	OOMLM		OOMLN		MitoEM-R		MitoEM-H	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
3D U-Net [3]	86.07	96.86	87.48	**99.11**	92.37	93.80	90.47	**92.92**
Swin UNETR [4]	96.79	96.69	97.97	98.48	93.90	89.31	76.94	88.89
Cellpose-SAM [9]	96.79	96.10	98.81	98.05	95.93	84.01	90.35	81.33
μSAM [1]	96.23	95.63	98.63	88.58	95.83	89.94	93.66	85.08
SAM4EM [10]	96.30	**97.20**	98.79	94.54	95.08	**95.17**	91.68	90.67
CCT U-Net [8]	88.77	95.04	93.84	81.97	72.61	89.05	63.78	78.69
GCT U-Net [6]	96.83	93.40	92.37	95.03	95.15	83.05	93.66	60.99
CPS U-Net [2]	95.62	95.88	**99.58**	56.08	94.57	92.25	91.41	86.38
Spatial-SAM	**97.32**	95.71	99.03	97.26	**96.40**	92.62	**93.74**	86.82

Table S2. **Comparison of average surface distance and 95th-percentile Hausdorff distance on different datasets (nm).** The **best** and second-best results for each metric are highlighted (lower is better).

Method	OOMLM		OOMLN		MitoEM-R		MitoEM-H	
	Avg Dist	HD95	Avg Dist	HD95	Avg Dist	HD95	Avg Dist	HD95
3D U-Net [3]	437.94	2651.61	961.33	8886.12	127.66	1086.18	88.84	955.18
Swin UNETR [4]	**13.49**	**25.26**	145.70	1506.97	49.60	578.44	186.15	1392.11
Cellpose-SAM [9]	23.71	77.79	**36.44**	**65.51**	100.93	661.20	109.17	729.53
μSAM [1]	23.29	132.21	214.69	2771.11	19.03	131.28	33.50	381.12
SAM4EM [10]	33.79	455.76	107.72	370.61	31.84	307.64	36.18	430.87
CCT U-Net [8]	197.29	1601.93	659.70	5384.88	263.27	1777.59	345.64	1671.54
GCT U-Net [6]	23.65	94.84	835.26	6640.49	32.06	167.30	66.21	532.50
CPS U-Net [2]	58.52	727.80	469.99	3552.18	31.20	321.06	46.18	521.85
Spatial-SAM	14.37	31.97	57.98	124.35	**14.55**	**47.97**	**29.12**	**259.46**

Table S3. **Comparison of surface Dice at 16nm on different datasets (%).** The **best** and second-best results for each metric are highlighted.

Method	OOMLM	OOMLN	MitoEM-R	MitoEM-H
3D U-Net [3]	80.14	56.76	85.08	**87.31**
Swin UNETR [4]	**94.55**	54.78	82.08	71.40
Cellpose-SAM [9]	92.89	54.88	74.92	72.76
μSAM [1]	88.77	38.65	85.43	81.40
SAM4EM [10]	92.23	36.02	**91.05**	84.86
CCT U-Net [8]	69.00	16.13	51.81	44.44
GCT U-Net [6]	83.87	21.02	72.75	57.90
CPS U-Net [2]	86.78	9.74	85.30	78.85
Spatial-SAM	92.03	**56.79**	90.42	85.75

is more robust than the classical maximum Hausdorff distance. Lower values indicate more accurate and less erratic boundaries. Across all evaluated datasets, Spatial-SAM ranks among the top two methods for both average surface distance and HD95. Notably, it achieves the best surface distances on MitoEM-R and MitoEM-H, and remains highly competitive on OpenOrganelle subsets (second-best on OOMLM and OOMLN). For example, on MitoEM-R it reduces the average surface distance from 19.03 nm (best baseline, μSAM) to 14.55 nm and HD95 from 131.28 nm to 47.97 nm, corresponding to substantially tighter and more stable mitochondrial surfaces. On OOMLN, Spatial-SAM delivers a competitive average distance and a large reduction in HD95 relative to semi-supervised and 3D baselines, while approaching the strongest 2D SAM-based model.

We further report the surface Dice similarity coefficient [7] at a 16 nm tolerance in Table S3. In cases where some methods produce substantial long-range segmentation errors and/or much noise on certain datasets, average surface distance and HD95 may not fully capture boundary quality; therefore we additionally include the surface Dice to reflect practical boundary agreement. Spatial-SAM attains surface Dice values that are uniformly within 2.6 percentage points of the highest score

on every dataset (differences: 2.52 on OOMLM, 0.00 on OOMLN, 0.63 on MitoEM-R, 1.56 on MitoEM-H), evidencing consistently strong boundary alignment across domains. Simultaneously, it delivers large gains over all semi-supervised baselines: +23.03/+8.16/+5.25 (OOMLM), +40.66/+35.77/+47.05 (OOMLN), +38.61/+17.67/+5.12 (MitoEM-R), and +41.31/+27.85/+6.90 (MitoEM-H) percentage points versus CCT/GCT/CPS respectively. These results show that Spatial-SAM provides boundary performance effectively on par with state-of-the-art fully supervised baselines while markedly surpassing semi-supervised approaches under limited annotation.

Table S4. Comparison of inference efficiency and resource consumption across different methods (Time: seconds; GPU RAM: graphics processing unit memory, GB; RAM: system memory, GB).

Method	2D Patch	3D Patch	Time	GPU RAM	RAM
3D U-Net	–	128^3	73	2.73	11.02
3D U-Net*	–	144^3	109	5.30	15.47
μSAM	512^2	–	608	4.19	18.59
Cellpose-SAM	256^2	–	4783	4.15	12.60
U-Net	512^2	–	99	1.33	2.90
Swin UNETR	–	96^3	1842	13.35	47.85
SAM4EM	512^2	–	182	1.41	4.46
Spatial-SAM	1024^2	144^3	106	9.41	13.68

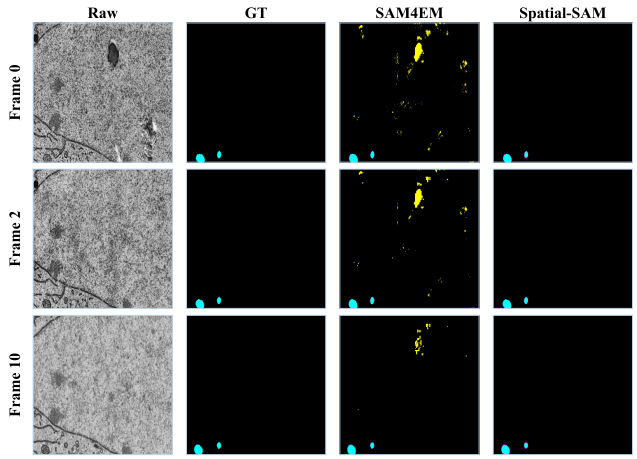

Figure S4. Visual comparison of error accumulation caused by sequential memory (SAM4EM) vs. geometry-aware SDF guidance (Spatial-SAM).

Performance Analysis. We evaluate the inference performance of different segmentation models on a volumetric electron microscopy dataset of size $1024 \times 1024 \times 1024$. All experiments are conducted on an NVIDIA RTX 3090 GPU. It is noted that different patch sizes, overlap and batch sizes could have a significant impact on the runtime and resource consumption, and our reported results serve as a representative example. For μSAM and Cellpose-SAM are tested using their default input resolutions as patch sizes (512^2 and 256^2, respectively). Our proposed method integrates both 2D and 3D pathways with patch sizes of 1024^2 and 144^3. Cellpose-SAM uses a batch size of 32, U-Net uses a batch size of 8, while other methods use a batch size of 1. The GPU RAM and RAM usages are measured as the peak consumption during the full inference of the volume.

In terms of efficiency, the 3D U-Net without overlap achieves the shortest runtime of 73.71 seconds due to the absence of redundant computation, whereas enabling overlap increases the runtime to 109.48 seconds and raises memory consumption accordingly. μSAM and Cellpose-SAM, both operating slice by slice in 2D, require substantially longer inference times of 608 seconds and 4783 seconds. Our method completes inference in 106 seconds, achieving a balance between 3D volumetric processing and 2D contextual efficiency. Compared with the SAM-based methods, our approach provides a significant speed advantage when scaled to volumetric data, demonstrating improved computational efficiency for large-scale 3D inference. Regarding training overhead, Spatial-SAM requires approximately 14.79 hours on MitoEM-R using a single NVIDIA RTX 3090.

4. Additional Discussion

Sequential Memory vs. SDF Guidance. While SAM2's stateful memory is intrinsically prone to error accumulation, SAM4EM exacerbates this issue by employing a momentum-updated feature memory. This momentum mechanism mathematically induces a delayed response, which results in slower memory updates compared to SAM2 and causes spatial misalignments during morphological changes. As shown in Fig. S4, when SAM4EM misclassifies an isolated slice-level artifact as a mitochondrion at Frame 0, the slow momentum update severely amplifies the error accumulation. This causes the false positive to linger and leave residual segmentations at Frame 10, long after the artifact has disappeared.

Spatial-SAM circumvents this issue by utilizing a geometry-driven signed distance field (SDF) memory. Because transient 2D artifacts rarely form coherent 3D structures across slices, the spatially continuous 3D SDF intrinsically acts as a structural filter. Consequently, Spatial-SAM not only suppresses the initial artifact misclassification at Frame 0 but also completely avoids the delayed reactions and error propagation seen in sequential memory mechanisms, demonstrating the robustness of explicit 3D spatial guidance.

Transferability and Adaptation Cost. Although the SAM2 module provides strong promptable priors, transferring to a new EM domain still typically requires fine-tuning or retraining due to the limited out-of-domain generalization of the U-Net SDF branch. To reduce adaptation over-

head, parameter-efficient fine-tuning (PEFT), such as LoRA and partial encoder freezing, represents a promising future direction. We expect such strategies to potentially maintain competitive segmentation quality while lowering both compute and memory requirements during model adaptation.

Applicability to More Complex Structures. The proposed pipeline is generally applicable to binary segmentation tasks beyond mitochondria. The SDF memory enforces geometry-aware cross-slice consistency, while the SAM2 module captures appearance variations. For objects with more complex topology or larger inter-slice deformation, we expect the same mechanism to remain beneficial, potentially with a larger memory neighborhood K and/or a moderately increased slice-level annotation ratio.

Multi-class Extension. For multi-class segmentation, a straightforward extension is to model class-wise SDFs in a multi-channel representation. A more compact alternative is to keep a shared foreground-background SDF memory for geometric guidance and add a semantic class head in the SAM2 branch for class discrimination. This design decouples geometric consistency from semantic categorization and can retain the efficiency advantages of the current framework.

References

[1] Anwai Archit, Luca Freckmann, Sushmita Nair, Nabeel Khalid, Paul Hilt, Vikas Rajashekar, Marei Freitag, Carolin Teuber, Melanie Spitzner, Constanza Tapia Contreras, et al. Segment anything for microscopy. *Nature Methods*, 22(3): 579–591, 2025. 1, 3

[2] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *CVPR*, pages 2613–2622, 2021. 1, 3

[3] 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*, pages 424–432. Springer, 2016. 1, 3

[4] 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 *MICCAI Brainlesion Workshop*, pages 272–284, 2021. 1, 3

[5] Larissa Heinrich, Davis Bennett, David Ackerman, Woohyun Park, John Bogovic, Nils Eckstein, Alyson Petruncio, Jody Clements, Song Pang, C Shan Xu, et al. Whole-cell organelle segmentation in volume electron microscopy. *Nature*, 599(7883):141–146, 2021. 1

[6] Zhanghan Ke, Di Qiu, Kaican Li, Qiong Yan, and Rynson WH Lau. Guided collaborative training for pixel-wise semi-supervised learning. In *ECCV*, pages 429–445, 2020. 1, 3

[7] Stanislav Nikolov, Sam Blackwell, Alexei Zverovitch, Ruheena Mendes, Michelle Livne, Jeffrey De Fauw, Yojan Patel, Clemens Meyer, Harry Askham, Bernardino Romera-Paredes, et al. Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. *arXiv preprint arXiv:1809.04430*, 2018. 3

[8] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *CVPR*, pages 12674–12684, 2020. 1, 3

[9] Marius Pachitariu, Michael Rariden, and Carsen Stringer. Cellpose-sam: superhuman generalization for cellular segmentation. *bioRxiv*, pages 2025–04, 2025. 1, 3

[10] Uzair Shah et al. Sam4em: Efficient memory-based two stage prompt-free segment anything model adapter for complex 3d neuroscience electron microscopy stacks. In *CVPR*. 3

[11] Abdel Aziz Taha and Allan Hanbury. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC medical imaging*, 15(1):29, 2015. 2

[12] Donglai Wei, Zudi Lin, Daniel Franco-Barranco, Nils Wendt, Xingyu Liu, Wenjie Yin, Xin Huang, Aarush Gupta, Won-Dong Jang, Xueying Wang, et al. Mitoem dataset: large-scale 3d mitochondria instance segmentation from em images. In *MICCAI*, pages 66–76, 2020. 1

[13] 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, et al. An open-access volume electron microscopy atlas of whole cells and tissues. *Nature*, 599(7883):147–151, 2021. 1