NeuroSeg Meets DINOv3: Transferring 2D Self-Supervised Visual Priors to 3D Neuron Segmentation via DINOv3 Initialization

Supplementary Material

This supplementary material provides additional technical details and extended results supporting the main paper. Specifically, we include: (i) definitions of the Dice and Cross-Entropy losses used in training, presented in Section 6; (ii) data preprocessing and augmentation procedures, described in Section 7; (iii) formal descriptions of segmentation and reconstruction metrics, detailed in Section 8; (iv) additional implementation details, including the modification of the DINOv3 downsampling stem, and the hyperparameter settings of TASL, provided in Section 9; (v) full quantitative results of additional baselines (TransUNet and SegMamba), reported in Section 10; (vi) cross-method evaluation of TASL on additional segmentation backbones in Section 11; (vii) computational analysis, indicated in Section 14; (viii) an analysis of pretraining effects on training effectiveness and feature behavior in Section 12; (ix) component-wise ablation of TASL in Section 13; and (x) additional qualitative visualizations across four neuron datasets, shown in Section 15.

6. Additional Loss Definitions

Following the main paper, our total objective combines Dice and Cross-Entropy (CE) losses. Below we provide the definitions of the Dice and CE losses.

Dice Loss. The Dice loss is a function that measures the overlap between the prediction and the ground truth. It is defined as:

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\sum_i p_i g_i}{\sum_i p_i^2 + \sum_i g_i^2 + \epsilon}, \tag{12}$$

where p_i and g_i represent the predicted probability and ground-truth label at voxel i, and ϵ is a small constant used to prevent division by zero.

Cross-Entropy Loss. The Cross-Entropy loss is employed to ensure voxel-wise classification accuracy:

$$\mathcal{L}_{\text{CE}} = -\sum_i g_i \log p_i + (1 - g_i)\log(1 - p_i). \tag{13}$$

7. Data Preprocessing and Augmentation

To ensure robust training and good generalization, we apply the following data preprocessing and augmentation steps:

- **Intensity Normalization:** Voxel intensities are normalized to zero mean and unit variance to reduce variability.
- **Patch-Based Cropping:** Due to GPU memory constraints, we adopt a patch-based strategy by randomly extracting sub-volumes such as $32 \times 32 \times 32$.
- **Data Augmentation:** To mitigate overfitting, we apply augmentations including random rotation, flipping, scaling, cropping, elastic deformation, gamma correction, brightness/contrast jitter, and Gaussian noise injection.

8. Evaluation Metric Definitions

Segmentation Metrics. Voxel-wise segmentation performance is evaluated using the F1 score and the 95th percentile Hausdorff Distance (HD95). A higher F1 score indicates better foreground prediction, while a lower HD95 reflects smaller boundary deviation between the prediction and the ground truth.

The F1 score is defined as:

$$\text{F1 Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{14}$$

where the Precision measures the proportion of correctly predicted foreground voxels among all voxels predicted as foreground, and Recall quantifies the proportion of actual foreground voxels that are correctly predicted.

The definition of HD95 is presented as follows:

$$\text{HD95} = \text{P}_{95}\left(\min_{y \in \mathcal{V}_{\text{gt}}} \|x - y\|\right), \quad x \in \mathcal{V}_{\text{pred}}, \tag{15}$$

where $\mathcal{V}_{\text{pred}}$ and $\mathcal{V}_{\text{gt}}$ denote the surface voxel sets of the prediction and the ground truth, respectively, $\|\cdot\|$ is the Euclidean distance, and $P_{95}(\cdot)$ extracts the 95th percentile.

Reconstruction Metrics. We adopt three standard tracing metrics widely used in neuron reconstruction: Entire Structure Average (ESA), Different Structure Average (DSA), and Percentage of Different Structure (PDS). Lower values of all three metrics indicate better topological consistency and reconstruction accuracy.

Given the predicted skeleton node set $\mathcal{S}_{\text{pred}}$ and the ground-truth node set $\mathcal{S}_{\text{gt}}$, ESA measures the average minimum distance from each node in the prediction to its closest node in the ground truth:

$$\text{ESA} = \frac{1}{|\mathcal{S}_{\text{pred}}|} \sum_{x \in \mathcal{S}_{\text{pred}}} \min_{y \in \mathcal{S}_{\text{gt}}} \|x - y\|. \tag{16}$$

DSA measures the mean distance between non-overlapping regions:

$$\text{DSA} = \frac{1}{|\mathcal{S}'_{\text{pred}}|} \sum_{x \in \mathcal{S}'_{\text{pred}}} \min_{y \in \mathcal{S}'_{\text{gt}}} \|x - y\|, \tag{17}$$

Table 6. **Quantitative comparison of neuron segmentation performance.** Each cell reports F1-score (%) / HD95 (lower is better for HD95) across four datasets. Columns correspond to datasets and rows to different segmentation methods. NeurINO^T and NeurINO^S denote the Tiny and Small variants of NeurINO. Best results are highlighted in **red**, and the second best in **blue**.

Method	Source	Params (M)	Drosophila	Mouse	NeuroFly	CWMBS
nnUNet [23]	*Nat. Methods 2021*	27.66	47.20 / 3.20	52.05 / 10.12	63.36 / 18.33	36.50 / 16.34
MedNeXt [47]	*MICCAI 2023*	61.97	47.74 / 3.15	50.61 / 13.77	62.50 / 19.23	33.46 / 18.37
TransUNet [8]	*MIA 2024*	105.28	34.88 / 16.67	38.43 / 48.13	37.93 / 18.55	24.56 / 21.14
SegMamba [67]	*MICCAI 2024*	67.42	37.48 / 5.03	39.98 / 56.37	46.77 / 17.12	28.01 / 19.57
NeurINO^T	-	39.21	**50.06 / 3.07**	**52.50 / 9.50**	**65.23 / 16.38**	**36.77 / 16.10**
NeurINO^S	-	61.52	**50.19 / 3.02**	**52.73 / 9.24**	**65.44 / 16.53**	**36.55 / 16.27**

Table 7. **Quantitative comparison of neuron tracing performance.** Each cell reports ESA / DSA / PDS (lower is better for all metrics) for two tracing algorithms: SmartTracing and NeuTube, evaluated across four datasets. NeurINO^T and NeurINO^S denote the Tiny and Small variants of NeurINO. Best results are highlighted in **red**, and the second best in **blue**.

Method	Drosophila		Mouse		NeuroFly		CWMBS	
	SmartTracing	*NeuTube*	*SmartTracing*	*NeuTube*	*SmartTracing*	*NeuTube*	*SmartTracing*	*NeuTube*
nnUNet [23]	1.67 / 4.48 / **0.20**	1.87 / 4.67 / 0.24	2.99 / 8.06 / **0.22**	5.36 / 16.10 / **0.22**	**22.78** / 31.76 / 0.41	**4.22** / 14.13 / **0.15**	36.93 / 42.10 / **0.53**	8.02 / 14.73 / **0.41**
MedNeXt [47]	1.68 / 4.44 / **0.20**	1.90 / 4.61 / 0.23	4.06 / 11.66 / 0.25	6.81 / 19.61 / **0.22**	27.52 / 34.74 / 0.40	4.32 / 14.06 / 0.17	34.74 / 38.94 / 0.56	7.93 / **13.58** / 0.44
TransUNet [8]	20.88 / 22.97 / 0.48	4.59 / 9.11 / 0.38	25.99 / 33.25 / 0.64	22.62 / 32.92 / 0.61	27.26 / 34.84 / 0.57	11.11 / 17.24 / 0.39	38.40 / 43.14 / 0.61	13.62 / 20.55 / 0.44
SegMamba [67]	43.58 / 46.37 / 0.46	5.14 / 8.23 / 0.32	30.28 / 41.19 / 0.55	35.41 / 49.75 / 0.55	22.82 / 31.53 / 0.47	8.97 / 18.53 / 0.29	37.54 / 42.56 / 0.57	10.57 / 17.70 / **0.40**
NeurINO^T	**1.62 / 4.29 / 0.20**	**1.75 / 4.18 / 0.21**	**2.81 / 7.88 / 0.22**	**4.86 / 15.30 / 0.22**	**21.93 / 28.98 / 0.32**	**4.13 / 13.91 / 0.16**	**34.09 / 37.76 / 0.54**	**7.74 / 13.40 / 0.41**
NeurINO^S	**1.65 / 4.40 / 0.21**	**1.84 / 4.31 / 0.22**	**2.92 / 7.73 / 0.21**	**4.73 / 15.53 / 0.21**	23.61 / **30.84 / 0.36**	4.24 / **14.02 / 0.15**	**33.48 / 38.09 / 0.54**	**7.89** / 13.67 / **0.41**

where $\mathcal{S}'_{\text{pred}}$ and $\mathcal{S}'_{\text{gt}}$ represent the non-overlapping skeleton nodes of the prediction and the ground truth.

PDS measures the proportion of mismatched branches over the total branch length:

$$\text{PDS} = \frac{\text{Length of mismatched branches}}{\text{Total branch length}}. \tag{18}$$

9. Additional Implementation Details

Modification of the DINOv3 Downsampling Stem. The original DINOv3 ConvNeXt-style downsampling stem performs a $4\times$ spatial reduction (stride 4) in the first layer. While suitable for 2D natural images, such aggressive resolution reduction tends to eliminate fine neuronal structures that are only a few voxels wide. To better preserve high-frequency structural cues critical for neurite continuity, we replace the stride-4 stem with a stride-2 variant and apply appropriate padding to maintain spatial alignment during 3D convolution, resulting in a $2\times$ downsampling at the input stage. This modification substantially improves the retention of thin neurites during early feature extraction and subsequently enhances both segmentation and reconstruction performance, as reported in Table 8.

Hyperparameters of TASL. For TASL, we use the weighting coefficients λ_{node}, λ_{edge}, and λ_{path} for the node-, edge-, and path-level terms, respectively. These coefficients are set to 1.0, 0.5, and 0.5. They are kept fixed in all experiments.

Table 8. **Effect of modifying the first downsampling ratio in DINOv3.** Better results are highlighted in **red**.

Downsampling Ratio	F1 (%)	HD95	*SmartTracing*			*NeuTube*		
			ESA	DSA	PDS	ESA	DSA	PDS
4×	47.08	4.86	1.80	**4.25**	0.24	1.98	4.47	0.25
2×	**50.06**	**3.07**	**1.62**	4.29	**0.20**	**1.75**	**4.18**	**0.21**

Table 9. **Cross-method evaluation of TASL.** Best results are highlighted in **bold**.

Method	F1 (%)	HD95	*SmartTracing*			*NeuTube*		
			ESA	DSA	PDS	ESA	DSA	PDS
nnUNet	**47.20**	3.20	1.67	4.48	**0.20**	1.87	4.67	0.24
nnUNet + TASL	47.09	**3.14**	**1.65**	**4.41**	0.21	**1.84**	**4.52**	**0.23**
MedNeXt	**47.74**	3.15	1.68	4.44	**0.20**	1.90	4.61	**0.23**
MedNeXt + TASL	47.66	**3.11**	**1.65**	**4.36**	0.21	**1.85**	**4.40**	0.24

10. Full Quantitative Results

Due to space limitations, the main paper reports a subset of baseline models. Here we provide the complete quantitative results, including additional baselines TransUNet [8] and SegMamba [67], as illustrated in Table 6 and Table 7.

11. Cross-method Evaluation of TASL

To evaluate the generality of TASL beyond the proposed NeurINO architecture, we integrate TASL into two representative 3D segmentation backbones, nnUNet and MedNeXt. Table 9 reports the results. TASL consistently im-

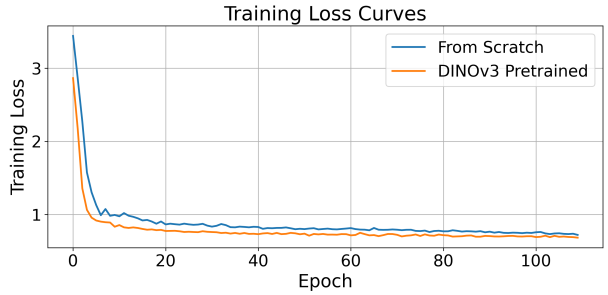

Figure 7. **Training loss curves comparing models trained from scratch and initialized with DINOv3 pretrained weights.** Pretraining accelerates convergence and stabilizes training dynamics.

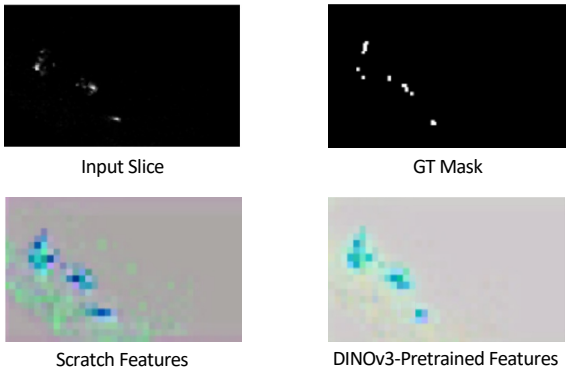

Figure 8. **Visualization of encoder features with and without DINOv3 pretraining.** Compared with features learned from scratch, DINOv3-pretrained features exhibit more coherent responses and clearer separation from the background.

proves topology-sensitive reconstruction metrics, demonstrating that its benefits are not limited to our architecture.

12. Effect of Pretraining on Training Effectiveness

Figure 7 compares training curves when initializing with DINOv3 pretrained weights versus training the model entirely from scratch. It is demonstrated that pretrained initialization leads to faster convergence, lower training loss, and smoother optimization, particularly in early training epochs.

To further analyze the effect of DINOv3 pretraining, we visualize encoder feature representations learned with and without pretraining. Given a 3D volume, we select a representative slice containing neuronal structures and extract encoder features. The feature maps are projected to three dimensions using principal component analysis (PCA) and visualized as RGB images. As shown in Figure 8, features learned from scratch tend to exhibit weaker structural coherence and noisier responses. In contrast, DINOv3-pretrained features produce more consistent activations along neuronal branches and better separation from the background. These results suggest that pretrained representations provide stronger intra-slice semantic cues, facilitating more effective feature aggregation across slices during training.

13. Component-wise Ablation of TASL

To further evaluate the contribution of each term in TASL, we conduct a component-wise ablation by removing different combinations of the node-, edge-, and path-level terms. Table 10 reports the results. Using all three components together yields the most balanced performance across segmentation and reconstruction metrics, confirming their complementary roles in preserving neuronal morphology.

Table 10. **Ablation of TASL components.** Best results are highlighted in **red**, and the second best in **blue**.

Variant	F1 (%)	HD95	*SmartTracing*			*NeuTube*		
			ESA	DSA	PDS	ESA	DSA	PDS
TASL (Node + Edge)	**49.94**	3.22	1.67	4.36	0.22	1.86	4.24	**0.21**
TASL (Node + Path)	49.86	**3.13**	**1.64**	**4.26**	0.22	1.83	**4.21**	0.22
TASL (Edge + Path)	49.75	3.20	1.66	4.34	**0.21**	**1.82**	4.29	0.22
TASL (Full)	**50.06**	**3.07**	**1.62**	**4.29**	**0.20**	**1.75**	**4.18**	**0.21**

14. Computational Analysis

We analyze the computational complexity of NeurINO and compare it with representative 3D segmentation baselines. Table 11 reports the number of parameters and floating-point operations (FLOPs) for each model. All FLOPs are measured for a single forward pass with input size $64 \times 64 \times 64$. Compared with conventional convolutional architectures such as nnUNet and MedNeXt, NeurINO achieves competitive or lower computational cost. In particular, NeurINO^T requires significantly fewer FLOPs while maintaining strong reconstruction performance, demonstrating its efficiency. Even the larger NeurINO^S variant has comparable computational complexity to MedNeXt while providing improved reconstruction quality.

Table 11. **Computational analysis of different models.** We report the number of parameters and FLOPs for representative segmentation methods.

Model	Params (M)	FLOPs (G)
nnUNet	27.66	55.68
MedNeXt	61.97	61.55
NeurINO^T	39.21	43.56
NeurINO^S	61.52	54.97

15. Additional Qualitative Visualizations

We provide additional visualization results across four datasets in Figures 9-12. Our method yields morphologically continuous reconstructions with fewer broken branches and more faithful neuronal arbor topology.

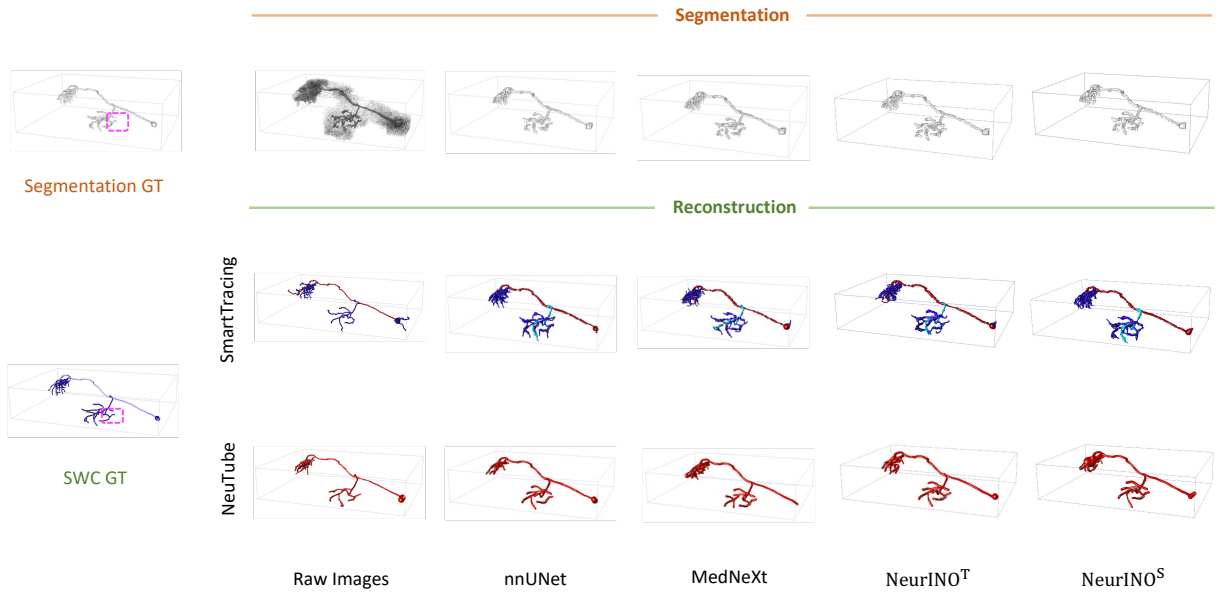

Figure 9. **Visualization comparison of the segmentation and tracing results on the Drosophila dataset.** The top row presents raw images and segmentation outputs, followed by two rows showing the corresponding reconstruction results generated by SmartTracing and NeuTube. **Magenta boxes** indicate severe false negatives (missed neurites) in other methods. Best viewed in zoom-in regions.

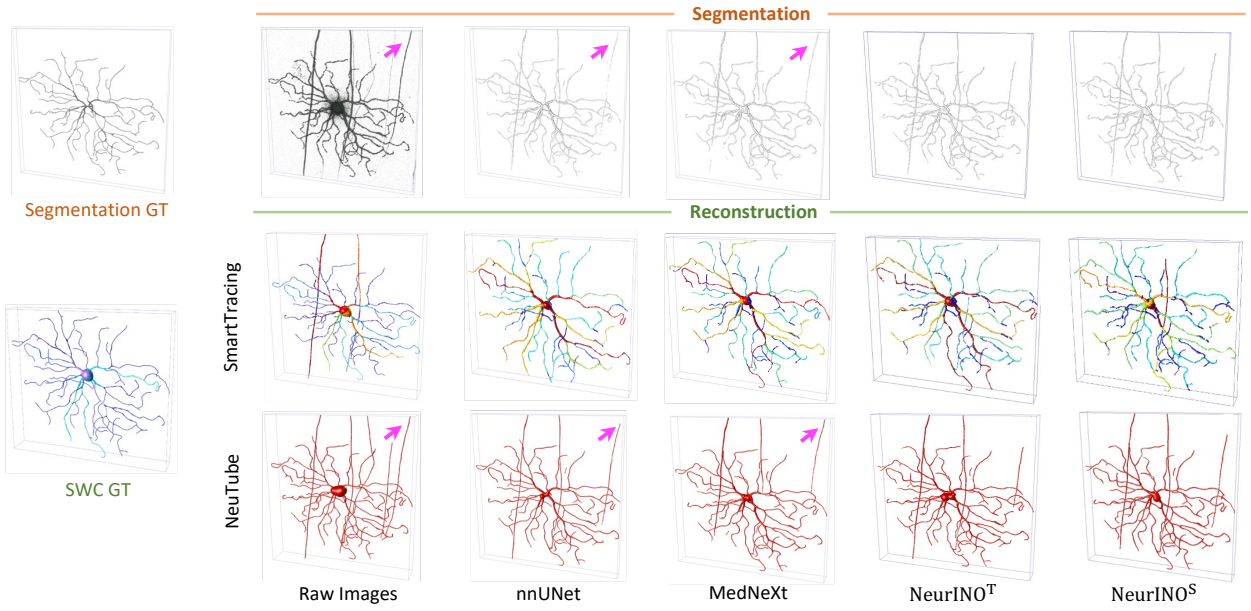

Figure 10. **Visualization comparison of the segmentation and tracing results on the Mouse dataset.** The top row presents raw images and segmentation outputs, followed by two rows showing the corresponding reconstruction results generated by SmartTracing and NeuTube. **Magenta arrows** highlight severe false positives in other methods. Best viewed in zoom-in regions.

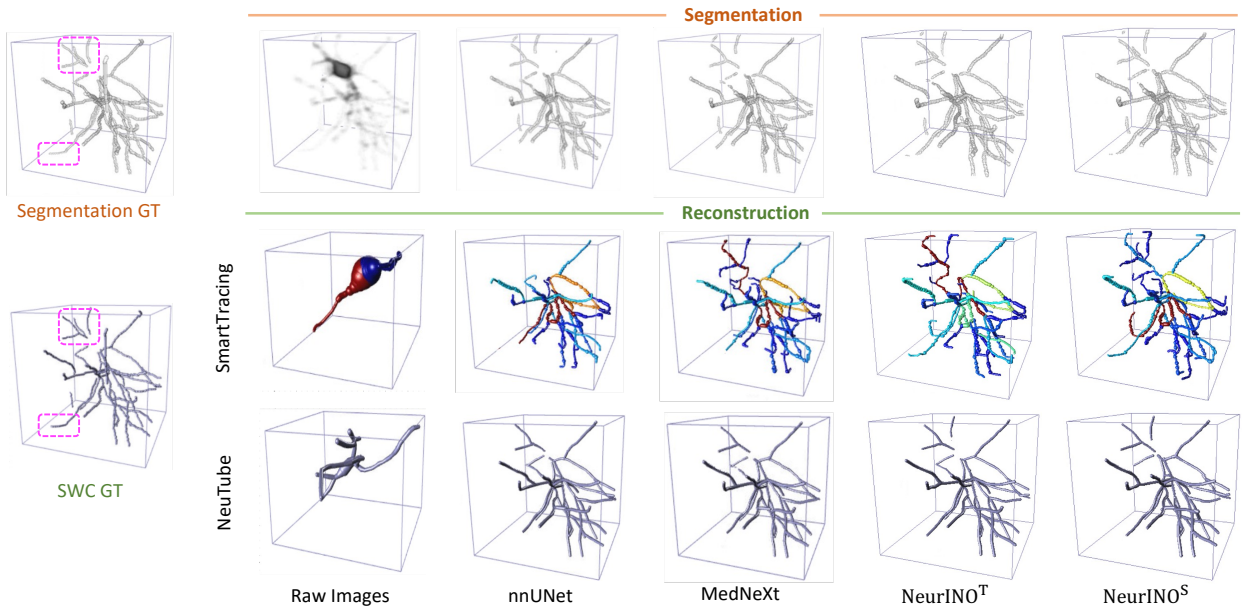

Figure 11. **Visualization comparison of the segmentation and tracing results on the NeuroFly dataset.** The top row presents raw images and segmentation outputs, followed by two rows showing the corresponding reconstruction results generated by SmartTracing and NeuTube. **Magenta boxes** indicate severe false negatives (missed neurites) in other methods. Best viewed in zoom-in regions.

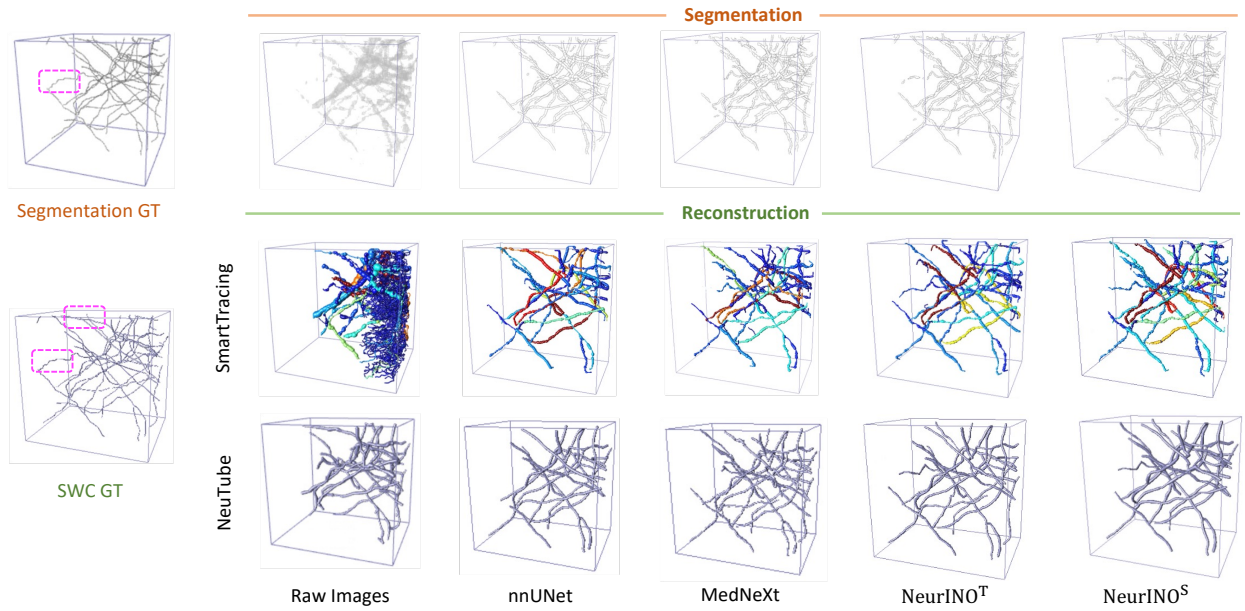

Figure 12. **Visualization comparison of the segmentation and tracing results on the CWMBS dataset.** The top row presents raw images and segmentation outputs, followed by two rows showing the corresponding reconstruction results generated by SmartTracing and NeuTube. **Magenta boxes** indicate severe false negatives (missed neurites) in other methods. Best viewed in zoom-in regions.