FGNet: Leveraging Feature-Guided Attention to Refine SAM2 for 3D EM Neuron Segmentation

Zhenghua Li[1, 2], **Hang Chen**[1, 2], **Zihao Sun**[3], **Kai Li**[1, 2], **Xiaolin Hu**[1, 2, 4*]

[1]Department of Computer Science and Technology, Institute for AI, BNRist, Tsinghua University, Beijing 100084, China
[2]Tsinghua Laboratory of Brain and Intelligence (THBI), IDG/McGovern Institute for Brain Research, Tsinghua University, Beijing 100084, China
[3]Zhili College, Tsinghua University, Beijing 100084, China
[4]Chinese Institute for Brain Research (CIBR), Beijing 100010, China
{li-zh24, chenhang20, szh24, li-k24}@mails.tsinghua.edu.cn, xlhu@tsinghua.edu.cn

Abstract

Accurate segmentation of neural structures in Electron Microscopy (EM) images is paramount for neuroscience. However, this task is challenged by intricate morphologies, low signal-to-noise ratios, and scarce annotations, limiting the accuracy and generalization of existing methods. To address these challenges, we seek to leverage the priors learned by visual foundation models on a vast amount of natural images to better tackle this task. Specifically, we propose a novel framework that can effectively transfer knowledge from Segment Anything 2 (SAM2), which is pre-trained on natural images, to the EM domain. We first use SAM2 to extract powerful, general-purpose features. To bridge the domain gap, we introduce a Feature-Guided Attention module that leverages semantic cues from SAM2 to guide a lightweight encoder, the Fine-Grained Encoder (FGE), in focusing on these challenging regions. Finally, a dual-affinity decoder generates both coarse and refined affinity maps. Experimental results demonstrate that our method achieves performance comparable to state-of-the-art (SOTA) approaches with the SAM2 weights frozen. Upon further fine-tuning on EM data, our method significantly outperforms existing SOTA methods. This study validates that transferring representations pre-trained on natural images, when combined with targeted domain-adaptive guidance, can effectively address the specific challenges in neuron segmentation.

Code — https://github.com/kwinderic/FGNet
Extended version — https://arxiv.org/pdf/2511.13063

1 Introduction

Neuron segmentation in electron microscopy (EM) images is important to neuroscience research, as accurate segmentation of complex neural structures is crucial for understanding brain connectivity and function (Funke et al. 2016; Kasthuri et al. 2015; Sheridan et al. 2023). However, this task remains highly challenging due to the inherent characteristics of EM data, 3D EM neuron images typically exhibit intricate morphological structures, low signal-to-noise ratios, and massive neuron populations, making it difficult to capture fine-grained details and maintain segmentation consistency (Lee et al. 2017).

Existing approaches to EM neuron segmentation can be roughly divided into two categories: (1) *Methods without pre-training*, which are trained from scratch on datasets specifically tailored for EM segmentation, including PEA (Huang et al. 2022), LSD (Sheridan et al. 2023), and CAD (Liu et al. 2024). (2) *Methods with EM-specific pre-training*, which aim to enhance generalization through pre-training strategies, including typically self-supervised learning such as DbMIM (Chen et al. 2023b) and EM-mamba (Chen et al. 2025b). However, all the aforementioned models are trained exclusively on EM data, which inherently limits their performance due to the scarcity of labeled EM datasets, as annotating complex 3D neural structures is extraordinarily difficult and labor-intensive.

Meanwhile, with the rise of pre-trained foundation models, significant progress has been made in the field of natural image segmentation. Models such as Mask2Former (Cheng et al. 2022), SAM (Kirillov et al. 2023), and SAM2 (Ravi et al. 2025), trained on large-scale natural image datasets with abundant segmentation masks, have demonstrated exceptional feature extraction capabilities and strong generalization abilities. *This raises a question: can the powerful representations learned from natural images be transferred to EM neuron segmentation to alleviate the data scarcity issue in the EM domain?*

In this paper, we try to answer this question. Considering that SAM2 is currently the latest and most widely used segmentation model, pre-trained on a large number of datasets, we mainly conducted our experiments on it. Previous work in medical image segmentation has utilized foundation models by fine-tuning or injecting adapters (Chen et al. 2023a). However, with these techniques we found that the results were not satisfactory (see Section 4.5). This challenge stems from the highly complex nature of neuronal image details. Architectures like SAM2, which incorporate multiple downsampling layers in their design, may inadvertently compromise some fine-grained structural information during image processing. *This observation highlighted the need for a dedicated refinement stage to recover lost details.*

To better incorporate fine-grained information by extracting details from the original image, existing works such

*Corresponding author.

as SAM-REF (Yu et al. 2025) adopt a strategy combining global and local refinement, targeting regions with high error rates, while BPR (Tang et al. 2021) employs boundary patch refinement, focusing specifically on boundary regions. Yet these strategies have a critical limitation in EM neuron segmentation: due to the large number of neuron instances in EM images, it is difficult to determine which specific regions need refinement in advance, making such methods unsuitable for this task. *This limitation motivates us to explore the possibility of refining the results directly using a separate, specialized network.*

Consequently, we design a lightweight network for refinement. To ensure effective results with minimal parameters, we develop a feature-guided attention module. This module guides the refinement network by leveraging features from the foundation model. Its purpose is to direct the network to extract fine-grained information from the original image, specifically the details that the foundation model may have overlooked during its downsampling stages.

Extensive experiments demonstrated that our methods achieved competitive performance even when SAM2's weights were frozen. Furthermore, with moderate fine-tuning on EM data, the proposed method outperformed state-of-the-art methods, providing a practical solution for balancing transfer efficiency and segmentation accuracy.

2 Related Work

2.1 Electron Microscopy Neuron Segmentation

Most methods involve designing expert models trained from scratch on specific datasets. The dominant approach predicts pixel affinities (encoding connectivity), followed by the watershed algorithm for fragment generation and aggregation to form complete neurons. Related improvements include PEA (Huang et al. 2022) using contrastive learning to enhance the robustness of affinities, and CAD (Liu et al. 2024) distilling 3D knowledge into 2D networks for efficiency. FragViT (Luo et al. 2024) utilizes a vision Transformer with hierarchical aggregation, and APViT (Sun et al. 2023) has a refinement module with appearance prompts. Other expert models within this paradigm include FFN (Januszewski et al. 2018), which segments by iteratively growing neuron fragments but is two orders of magnitude slower than affinity-based methods (Chen et al. 2025a); AGQ designs affinity-guided query initialization combined with learnable queries to achieve end-to-end segmentation but requires adjusting the number of queries for specific data.

Beyond expert models, EM-specific pre-training strategies have emerged, such as self-supervised pre-training in DbMIM (Chen et al. 2023b) and long-range modeling in EMmamba (Chen et al. 2025b), which can improve generalization ability but suffer from limited EM data.

2.2 Adaptation from Foundation Models

Foundation models like Segment Anything Model (SAM) (Kirillov et al. 2023) and its successor SAM2 (Ravi et al. 2025) have demonstrated remarkable capabilities, trained on large-scale datasets and exhibiting strong zero-shot performance across diverse segmentation tasks (Roy et al. 2023). However, despite their strong zero-shot capabilities, they still underperform on specific downstream tasks. MedSAM (Ma et al. 2024) adapts SAM to medical segmentation by freezing the prompt encoder and fine-tuning the image encoder/mask decoder for domain alignment. Medical SAM Adapter (Wu et al. 2025) uses lightweight adapters for parameter-efficient adaptation without full fine-tuning. SemiSAM (Zhang et al. 2024b) leverages SAM as an auxiliary supervision branch, generating pseudo-labels for semi-supervised learning. SAM-Med3D (Wang et al. 2024) extends SAM to 3D by converting core components for volumetric medical data. SAM-REF (Yu et al. 2025) improves quality via global/local refinement. These methods demonstrate various strategies for adapting and refining foundation models for specialized tasks.

2.3 Attention Modules in Segmentation and Detection

Attention mechanisms enhance feature representation by focusing on critical regions, with diverse modules tailored for segmentation and detection: Woo et al. (Woo et al. 2018) proposed CBAM, a lightweight module that infers channel and spatial attention maps sequentially; Hu et al. (Hu, Shen, and Sun 2018) introduced SENet, a pioneering channel attention framework modeling inter-channel dependencies via global pooling and fully connected layers; Wang et al. (Wang et al. 2018) extended attention to global contexts with Non-local Networks, capturing long-range dependencies; Xu et al. (Xu et al. 2019) presented AC-FPN, which uses Context and Content Attention Modules to enhance multi-scale fusion; Fu et al. (Fu et al. 2019) designed DANet, combining position and channel attention to model global dependencies; Yang et al. (Yang et al. 2019) proposed DEA-Net with Content-Guided Attention; Wang et al. (Wang et al. 2020) developed ECA-Net, an efficient channel attention variant using 1D convolutions; Li et al. (Li et al. 2024) introduced VL-SAM, leveraging attention maps as prompts for open-vocabulary segmentation. These modules cover channel, spatial, and global attention, demonstrating versatility in feature refinement, and our work builds on this by using pre-trained weights to guide attention modules in extracting fine-grained details.

3 Methods

3.1 Overall Pipeline

Let the volume of electron microscopy (EM) neuron images be denoted as $V \in \mathcal{R}^{D\times H\times W}$, where D, H, and W represent the depth, height, and width dimensions, respectively. Our segmentation model predicts affinity matrices via a hierarchical pipeline (Figure 1(a)) with four core components: SAM2 Encoder, Feature-Guided Attention (FGA), Fine-Grained Encoder (FGE), and dual affinity decoders (shared architecture, distinct inputs). The SAM2 Encoder processes V to generate multi-scale features S, which are fed into the first decoder to produce a coarse affinity matrix A_{s}. FGA then modulates S to guide FGE in generating fine-grained features G, which the second decoder uses to derive

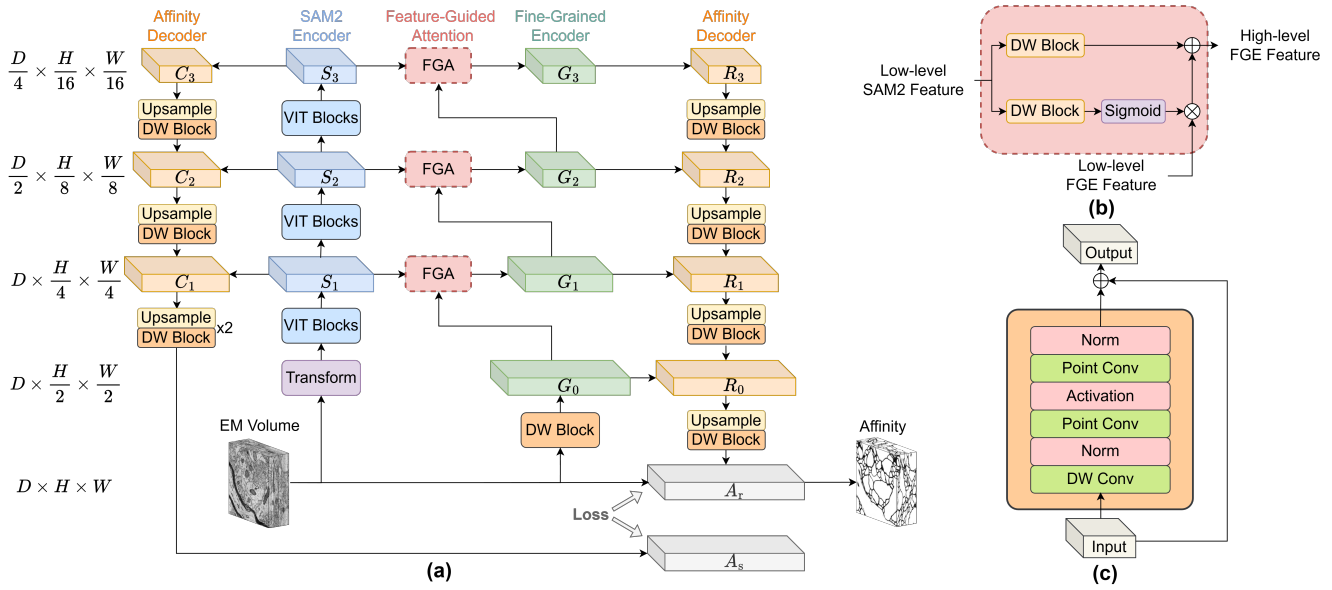

Figure 1: Network architecture. (a) Four core components: SAM2 Encoder, Feature-Guided Attention (FGA), Fine-Grained Encoder (FGE), and a pair of Affinity Decoders. Modules of different types are represented with distinct colors. The left column indicates the size of each feature in the horizontal dimension. (b) Detailed structure of the FGA module. The plus sign ($\oplus$) represents element-wise addition, and the multiplication sign ($\otimes$) represents element-wise multiplication. (c) Detailed structure of the DW Block.

a refined matrix A_{r}. Both A_{s} and A_{r} are supervised against A_{gt}.

During inference, the affinity decoder on the left in Figure 1(a) is not used, while affinity A_{r} is then processed using watershed (Funke et al. 2018) and agglomeration (Funke et al. 2018) algorithms to generate the final segmentation result $O \in \mathcal{R}^{D\times H\times W}$.

3.2 SAM2 Encoder

The input EM volume is first processed by a transformation module to align with SAM2's input specifications, followed by a sequence of ViT Blocks (Ravi et al. 2025) to generate hierarchical features S_i (with $i \in \{1, 2, 3\}$ indicating distinct levels). This component is visualized as the blue section in Figure 1(a).

3.3 FGA Module and Fine-Grained Encoder

The FGE extracts fine-grained features G_i (with $i \in \{0, 1, 2, 3\}$ indicating distinct levels) from the original EM volume V, guided by the SAM2 Encoder's hierarchical features S_i through the FGA module. Specifically, the initial fine-grained feature G_0 is derived directly from V as:

$$G_0 = \mathrm{DW}(V), \tag{1}$$

where $\mathrm{DW}(\cdot)$ denotes the depth-wise convolution block (abbreviated as DW Block), whose structure is depicted in Figure 1(c).

The FGA module refines subsequent features G_i from the preceding G_{i-1} under S_i guidance, employing a two-branch attention mechanism (Figure 1(b)). For each level i, SAM2 Feature S_i first generates two attention maps a_i and b_i to capture distinct structural priors:

$$a_i = \sigma\left(\mathrm{DW}(S_i)\right), \tag{2}$$

$$b_i = \mathrm{DW}(S_i), \tag{3}$$

where $\sigma(\cdot)$ is the sigmoid activation function constraining $a_i \in [0, 1]$. The refined feature G_i is computed by

$$G_i = a_i \odot G_{i-1} + b_i, \tag{4}$$

where $\odot$ denotes element-wise multiplication. This design enables a_i to weight informative regions in G_{i-1} and b_i to compensate for feature biases.

3.4 Affinity Decoder

We employ two affinity decoders with identical architecture but varying depths and inputs, as shown in Figure 1(a).

The left decoder processes the hierarchical features S_i from the SAM2 Encoder to predict the coarse affinity matrix A_{s}. It consists of a sequence of upsampling operations and DW Blocks, which progressively restore the feature resolution to match the input volume dimensions.

The right decoder, sharing the same structure design but with a different number of layers, takes the fine-grained features G_i as input. Through a parallel sequence of upsampling steps and DW Blocks, it generates the refined affinity matrix A_{r}. Both decoders leverage DW Blocks to preserve computational efficiency while maintaining computational efficiency, with their distinct layer counts tailored to the characteristics of their respective input features (S_i for coarse prediction and G_i for fine-grained refinement).

Notably, the left Affinity Decoder functions as a deep supervision branch. It provides supervisory signals to enable SAM2's feature extraction capabilities to adapt as much as possible to EM data. However, its output remains coarse due to its primary focus on initial feature adaptation. In contrast, the right decoder produces refined results by extracting fine-grained features under the guidance of the FGA module, thus serving as the final affinity result.

4 Experiments

4.1 Datasets and Metrics

We evaluated our method on three widely-used public electron microscopy (EM) neuron segmentation datasets:

- **AC3/AC4** (Kasthuri et al. 2015): Labeled subsets derived from the mouse somatosensory cortex dataset. The dataset images are acquired at a spatial resolution of $3 \times 3 \times 29$ nm^3. Specifically, the AC3 subset consists of 256 sequential sections, while the AC4 subset contains 100 sequential sections. The data partitioning follows the protocol in (Chen et al. 2025a): the first 80 sections of AC4 are allocated for training, the subsequent 20 sections of AC4 for validation, and the first 100 sections of AC3 for testing, with each section in both subsets having dimensions of 1024×1024 pixels.
- **CREMI** (Funke et al. 2016): A benchmark dataset containing 3D EM volumes of Drosophila melanogaster brain tissue, with a resolution of $4 \times 4 \times 40$ nm^3. The dataset includes 3 volumes (A, B, C) each with dimensions $1250 \times 1250 \times 125$, and provides manually annotated segmentation masks for training and evaluation. Every volume is divided into 100 sections for training and 25 sections for testing.
- **Wafer4** (Liu et al. 2024): Collected from a region of the mouse medial entorhinal cortex, which has a size of 1250×1250×125 voxels. This dataset is divided into 100 sections for training and 25 sections for testing.

We utilized two widely adopted metrics to quantitatively evaluate segmentation results, both of which measure error (where lower values indicate better performance):

- **Variation of Information (VOI)** : A dissimilarity metric that quantifies the discrepancy between predicted and ground truth segmentations(Meilă 2003), explicitly accounting for both over-segmentation and over-merging errors. It decomposes into two components: $\text{VOI}_{\text{split}}$ (assessing over-segmentation) and $\text{VOI}_{\text{merge}}$ (assessing over-merging).
- **Adapted Rand Error (Arand)** : An adjusted variant of the Rand Index(Rand 1971), optimized to address the uneven distribution of object sizes in EM image segmentation tasks. This metric measures the disagreement between predicted results and ground truth.

4.2 Implementation Details

The input block size for both training and inference was set to $16 \times 256 \times 256$, with a stride of $8 \times 128 \times 128$ used for inference. The model was optimized via the Adam optimizer (Kingma and Ba 2017) with an initial learning rate of 1×10^{-4}, training was conducted with a batch size of 4 over 10,000 iterations. Our framework was implemented using pytorch_connectomics codebase (Lin et al. 2021). Our experiments were conducted trained on 4 NVIDIA 3090 GPUs, each with 24GB memory. See more in the Appendix.

4.3 Quantitative Results

To thoroughly evaluate the effectiveness of the proposed method, we compared our method with state-of-the-art EM neuron segmentation methods on the AC3/AC4, CREMI and Wafer4 datasets, respectively. Our approach exhibited notable performance characteristics across different training configurations:

Our model achieved state-of-the-art (SOTA) performance across multiple datasets (see Tables 1, 2, and 3). Notably, on the AC3/AC4 datasets, our method demonstrated strong performance in the VOI metric compared to the previous SOTA method CAD (Liu et al. 2024). Notably, our model achieved performance comparable to CAD, even without fine-tuning the pre-trained SAM2 encoder. Furthermore, after fine-tuning, our approach achieved a substantial 12.5% improvement over CAD, solidifying its superior performance. This superior performance extended to other benchmarks: on CREMI A, B, and C datasets, our fine-tuned model outperformed existing SOTA methods by 0.7%, 1.0%, and 3.7% respectively in VOI score; on the Wafer4 dataset, we observed a notable 4.6% improvement in VOI score.

These results validated the effectiveness of our framework in balancing large-scale pre-training advantages with dataset-specific optimization, enabling consistent performance boosts across diverse EM imaging benchmarks.

4.4 Qualitative Results

We visualize the results of the proposed method in comparison with those of previous methods, showcasing both the 2D sections (In Figure 2) and the 3D neuron morphology (In Figure 3).

The 2D slices present pixel-level segmentation results. As highlighted by the orange boxes, previous methods exhibit numerous over-segmentation and under-segmentation errors, whereas our method successfully separates these regions. In the 3D morphology visualization, missegmented areas in previous methods are highlighted with red arrows, particularly where multiple fine-scale synapses of the main neuron were not accurately predicted. In contrast, our method achieves a more precise reconstruction of the neuron's structure.

4.5 Comparison of Foundation Model Adaptation Methods

As we mentioned in Section 1, we first compared different adaptation methods for foundation models. The following are the adaptation approaches for different foundation models such as SAM and SAM2, which we have implemented and applied to EM segmentation for comparison. Specific implementation details can be found in the Appendix. *Frozen*: Direct transfer with most pre-trained parameters frozen (Roy et al. 2023); *Fine-tuning*: Full fine-tuning

Model		Reference	$\text{VOI}_{\text{split}}$	$\text{VOI}_{\text{merge}}$	VOI	Arand
No Pre-training	ResUNet (Çiçek et al. 2016)	MICCAI'16	1.037*	0.258*	1.295*	0.154*
	SuperHuman (Lee et al. 2017)	ArXiv'17	1.145*	0.263*	1.408*	0.122*
	MALA (Funke et al. 2018)	TPAMI'19	1.304*	0.242*	1.546*	0.120*
	SEUNet (Lin et al. 2021)	ArXiv'21	1.031*	0.251*	1.282*	0.156*
	SwinUNETR (Hatamizadeh et al. 2021)	MICCAI'21	1.238*	0.191*	1.429*	0.110*
	PEA (Huang et al. 2022)	AAAI'22	0.852*	0.232*	1.084*	0.094*
	UNETR (Hatamizadeh et al. 2022)	WACV'22	1.048*	0.237*	1.285*	0.116*
	LSD (Sheridan et al. 2023)	NM'23	1.448*	0.229*	1.677*	0.134*
	APViT (Sun et al. 2023)	IJCAI'23	0.767	0.204	0.971	0.078
	FragViT (Luo et al. 2024)	AAAI'24	0.868	0.191	1.054	0.093
	CAD (Liu et al. 2024)	CVPR'24	0.601	0.431	1.032	0.119
	CAD + KD (Liu et al. 2024)	CVPR'24	0.533	0.351	0.884	0.081
	AGQ (Chen et al. 2025a)	ICLR'25	0.677	0.290	0.967	0.095
Pre-trained on EM	DbMIM+UNETR (Chen et al. 2023b)	IJCAI'23	0.647	0.285	0.931	0.243
	SegNeuron (Zhang et al. 2024a)	MICCAI'24	0.698*	0.245*	0.943*	0.088*
	EMmamba (Chen et al. 2025b)	ICCV'25	0.938	0.863	1.801	0.284
	EMmamba† (Chen et al. 2025b)	ICCV'25	0.448	0.544	0.992	0.137
Pre-trained on Natural Images	Ours w/o finetune SAM2	-	0.647	0.263	0.910	0.096
	Ours	-	0.614	0.183	**0.797**	**0.069**

Table 1: Comparison of different models on AC3/AC4 datasets. Lower values are better for all metrics. * indicates results are reproduced by us. † denotes a modified version by the original authors. VOI results are obtained by the Waterz (Funke et al. 2018) post-processing.

	CREMI-A				CREMI-B				CREMI-C			
Model	VOI_{s}	VOI_{m}	VOI	Arand	VOI_{s}	VOI_{m}	VOI	Arand	VOI_{s}	VOI_{m}	VOI	Arand
SuperHuman (Lee et al. 2017)	0.399	0.241	0.640	0.089	0.554	0.222	0.776	0.048	0.820	0.338	1.158	0.179
MALA (Funke et al. 2018)	0.398	0.236	0.634	0.085	0.589	0.261	0.850	0.041	0.842	0.332	1.174	0.162
PEA (Huang et al. 2022)	0.329	0.298	0.626	0.091	0.411	0.374	0.785	0.041	0.745	0.446	1.191	0.169
APViT (Sun et al. 2023)	0.445	0.260	0.704	0.117	0.579	0.201	0.781	0.032	0.884	0.234	1.118	0.110
DbMIM+UNETR (Chen et al. 2023b)	0.411	0.331	0.743	0.131	0.642	0.381	1.023	0.092	0.925	0.276	1.201	0.107
CAD (Liu et al. 2024)	0.326	0.299	0.625	0.107	0.402	0.347	0.749	0.045	0.738	0.455	1.193	0.170
CAD + KD (Liu et al. 2024)	0.313	0.252	0.565	0.079	0.379	0.305	0.684	**0.030**	0.738	0.322	1.060	0.149
Ours w/o finetune SAM2	0.393	0.250	0.643	0.101	0.398	0.380	0.778	0.061	0.901	0.299	1.201	0.149
Ours	0.331	0.230	**0.561**	**0.075**	0.514	0.163	**0.677**	0.040	0.540	0.481	**1.021**	**0.081**

Table 2: Results on the CREMI dataset. Results on three volumes are reported. Note that VOI is the overall metric of VOI_{s} and VOI_{m}. Results were borrowed from (Liu et al. 2024). The best and second-best results are bolded and underlined, respectively.

Model	VOI_{s}	VOI_{m}	VOI	Arand
SuperHuman	0.452	0.166	0.618	0.041
MALA	0.455	0.158	0.613	0.036
PEA	0.421	0.172	0.593	0.034
APViT	0.581	0.123	0.704	0.036
CAD	0.404	0.224	0.627	0.051
CAD + KD	0.415	0.144	0.559	0.030
Ours w/o finetune SAM2	0.598	0.131	0.729	0.042
Ours	0.412	0.121	**0.533**	**0.028**

Table 3: Results on the Wafer4 dataset. Results were obtained from (Liu et al. 2024). The best results are bolded.

of the foundation model (Ma et al. 2024); *Refinement*: Direct integration of a learnable lightweight refinement network (Yu et al. 2025).

Experimental results in Table 4 demonstrate that our method, by incorporating a feature-guided attention mechanism between the foundation model and the lightweight network, outperforms all the above approaches. This is primarily because the substantial domain gap between natural images and EM data, coupled with significant discrepancies in signal-to-noise ratio, renders both frozen and naively fine-tuned models ineffective at capturing fine-grained details. While adding an extra network yields better refinement, exclusive reliance on refinement without leveraging the general feature guidance from the foundation model overlooks its supervisory value in aligning with domain-agnostic patterns. These results further confirmed that the integration of both fine-tuning and an adaptive lightweight network (as enabled by our design) is essential for achieving optimal performance in EM neuron segmentation.

4.6 Feature-Guided Attention Analysis

To address SAM2's coarse feature extraction in EM segmentation, we designed FGA module to recover the overlooked fine details (see Figure 1(b)). To validate our feature-guided attention mechanism, we visualize attention maps (which refer to a_i in Equation 2) (Figure 4) from EM neuron segmentation. The visualization presents input EM slices, ground

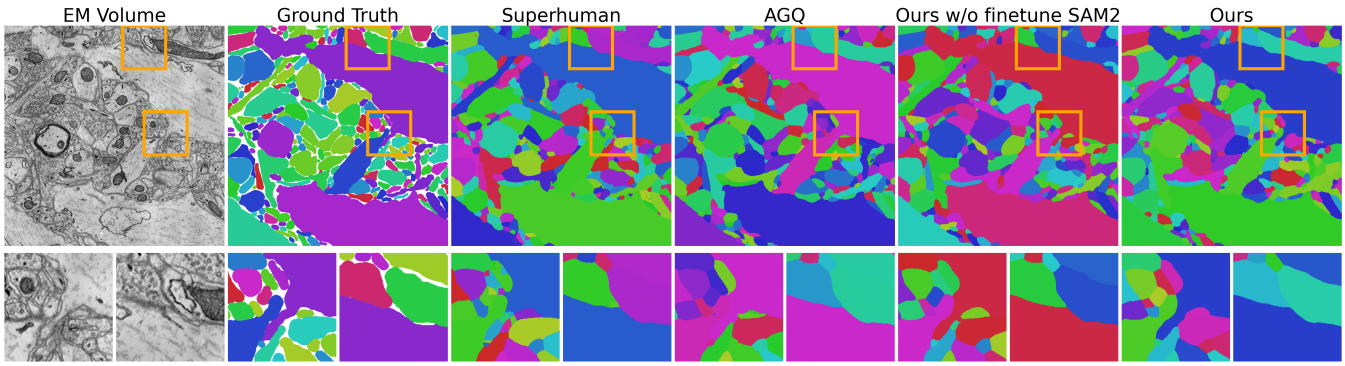

Figure 2: 2D visualization of segmentation results obtained from different methods. For each result, the first row presents a 2D slice and the second row shows two zoomed-in regions. Best viewed in digital with zoom-in.

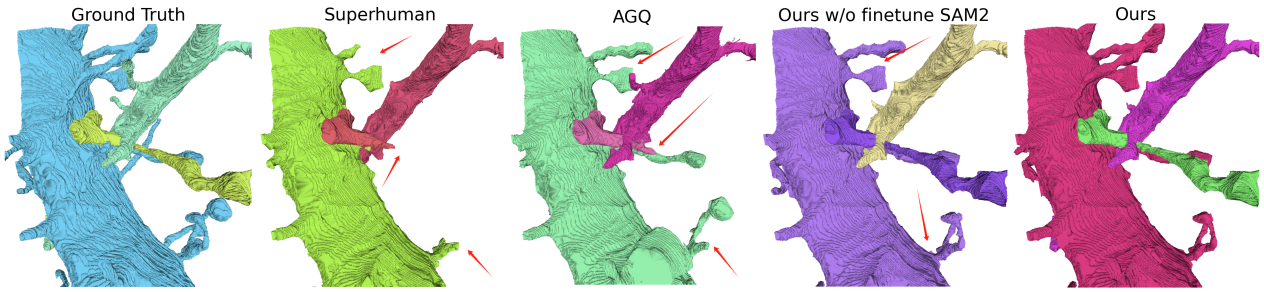

Figure 3: 3D visualization of segmentation results obtained from different methods. The 3D neuron morphology is shown, with red arrows indicating some of the slender dendrites in the neuron.

Adapting Methods	VOI_s	VOI_m	VOI	Arand
Frozen	1.318	0.332	1.649	0.147
Fine-tuning	1.176	0.179	1.355	0.108
Refinement	0.679	0.372	1.051	0.142
Our Method	0.614	0.183	**0.797**	**0.069**

Table 4: Comparison of different methods for adapting the foundation model. Lower values are better for all metrics.

Model Variants	VOI_s	VOI_m	VOI	Arand
w/o SAM2 Encoder	0.682	0.426	1.108	0.164
w/o FGE	1.176	0.179	1.355	0.108
w/o FGA Block	0.679	0.372	1.051	0.142
All Modules	0.614	0.183	**0.797**	**0.069**

Table 5: Ablation study on different modules. Lower values are better for all metrics.

FGA Type	VOI_s	VOI_m	VOI	Arand
SE-Type	0.671	0.308	0.979	0.110
CABM-Type	0.608	0.276	0.883	0.104
ECA-Type	0.604	0.235	0.839	0.091
Our Method	0.614	0.183	**0.797**	**0.069**

Table 6: Ablation study on different FGA module. Lower values are better for all metrics.

truth segmentations and hierarchical attention maps. A jet colormap color bar quantifies attention weights, where warm colors (red/yellow) indicate higher attention and cool colors (blue/green) represent lower weights.

Key observations confirm the mechanism's effectiveness: the attention maps exhibit a progressive refinement across a_1 to a_3, shifting focus from broader neurite regions to increasingly fine-grained structural details in alignment with neural structural hierarchy. Domain-specific focus is evident as high-attention regions consistently coincide with critical segmentation structures such as neuron boundaries, small neurites, and synapses, while lower attention weights are concentrated in homogeneous cell interiors and irrelevant organelles.

This discriminative attention pattern reflects EM-specific knowledge that the natural image-pretrained SAM2 encoder lacks. Our feature-guided attention mechanism enables FGE to capture these critical details, effectively addressing limitations in cross-domain transfer.

4.7 Ablation Study

We conducted a series of ablation experiments on the AC3/AC4 datasets to verify the effectiveness of our proposed method.

Ablation of Network Modules In Table 5, we further ablated different modules in the proposed network to evaluate their individual contributions. Removing any compo-

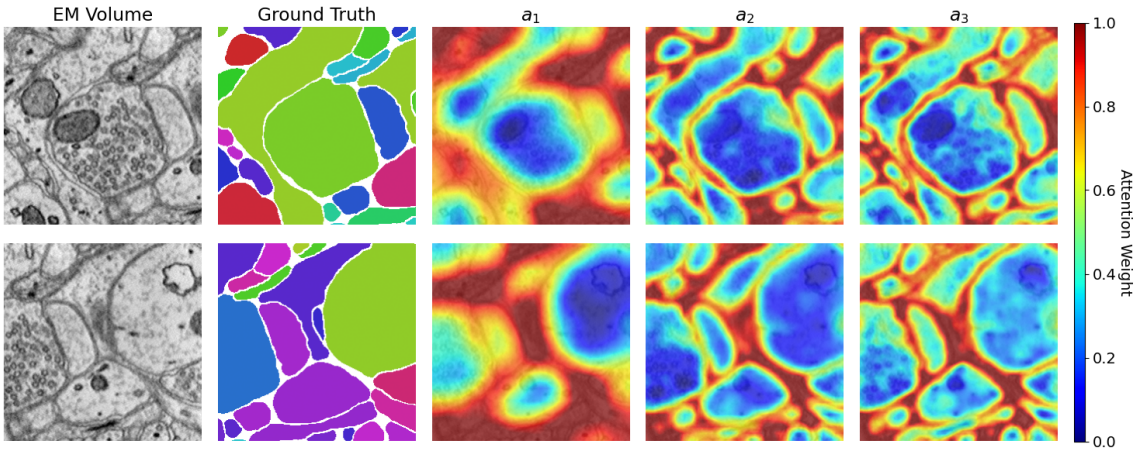

Figure 4: Hierarchical attention visualization for 3D EM neuron segmentation. Rows show sequential 2D slices. Columns 1–2 display input EM volumes and ground truth segmentations.Columns 3–5 display attention maps (a_1, a_2, a_3) overlaid on EM data, with warmer jet colormap colors indicating higher attention weights. A vertical color bar quantifies attention intensity, highlighting critical structural details like neuron boundaries.

Depth	VOI_s	VOI_m	VOI	Arand
3	0.624	0.192	0.816	0.090
4	0.614	0.183	**0.797**	**0.069**
5	0.608	0.276	0.883	0.104

Table 7: Ablation study on different depth of FGE. Lower values are better for all metrics.

nent, whether SAM2, the FGE, or the FGA, led to performance degradation, confirming the effectiveness of our proposed framework as an integrated paradigm. Specifically, the absence of FGE resulted in diminished fine-grained segmentation performance, as its refining capability is critical for capturing intricate structural details. Without the FGA, the guidance mechanism between SAM2 and FGE was disrupted, leading to suboptimal alignment of attention toward key regions. Meanwhile, removing SAM2 caused the loss of coarse feature guidance, which serves as the foundational structure for accurate segmentation. Together, these results underscore that each module plays an indispensable role in achieving the overall performance, validating the rationality of our integrated design.

Comparison of FGA Implementations In designing our FGA module, we drew inspiration from attention mechanisms proposed in prior works. Specifically, the SE-Type (Hu, Shen, and Sun 2018) Module achieves channel-wise attention by squeezing feature maps into global descriptors via average pooling and recalibrating channel weights through two fully connected layers. The CBAM-Type (Woo et al. 2018) Module extends this by incorporating both channel attention and spatial attention, which is computed via a convolution layer applied to concatenated average and maximum pooled features. The ECA-Type (Wang et al. 2020) Module simplifies channel attention by replacing fully connected layers with a 1D convolution, adapting its kernel size to the number of channels for more efficient recalibration. To validate the effectiveness of our specific design, we conducted comparative experiments with these existing attention mechanisms. Experimental results in Table 6 demonstrate that our designed attention mechanism outperforms alternatives, as it is uniquely tailored to our paradigm of guiding fine-grained feature extraction using coarse features. This superiority confirms the rationality of our FGA design, which is specifically optimized for the cross-scale guidance scenario in our framework.

Impact of FGE Depth Finally, we analyzed the impact of FGE depth in Table 7. As a lightweight refinement network, FGE does not require excessive parameters, and a shallow architecture with only a few layers suffices. Experimental results validate that increasing network depth has minimal effect on performance, confirming that deeper layers are unnecessary for effective refinement and further supporting the efficiency of our lightweight design. This is because foundation models like SAM2 already encode a wealth of generic and coarse-grained features, allowing the refinement network to focus exclusively on capturing specific fine-grained details. Consequently, FGE can achieve excellent performance with relatively few parameters.

5 Conclusion

We propose FGNet, a 3D EM neuron segmentation framework that transfers knowledge from SAM2 to the EM domain. By leveraging the powerful feature extraction capabilities of SAM2 via a Feature-Guided Attention module, our method effectively bridges the domain gap between natural images and 3D EM data. Experiments validated its competitive performance with frozen SAM2 weights and significant SOTA outperformance after fine-tuning.

Acknowledgements

This work was supported by the National Key Research and Development Program of China under Grant 2021ZD0200301, the National Natural Science Foundation of China under Grants 62576187 and U2341228.

References

Chen, H.; Tang, C.; Li, X.; and Hu, X. 2025a. Efficient Neuron Segmentation in Electron Microscopy by Affinity-Guided Queries. In *Proceedings of the Thirteenth International Conference on Learning Representations*.

Chen, T.; Zhu, L.; Deng, C.; Cao, R.; Wang, Y.; Zhang, S.; Li, Z.; Sun, L.; Zang, Y.; and Mao, P. 2023a. Sam-adapter: Adapting segment anything in underperformed scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3367–3375.

Chen, Y.; Huang, W.; Zhou, S.; Chen, Q.; and Xiong, Z. 2023b. Self-Supervised Neuron Segmentation with Multi-Agent Reinforcement Learning. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 609–617. ijcai.org.

Chen, Y.; Shi, H.; Liu, X.; Shi, T.; Zhang, R.; Liu, D.; Xiong, Z.; and Wu, F. 2025b. TokenUnify: Scaling Up Autoregressive Pretraining for Neuron Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13604–13613.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1290–1299.

Çiçek, Ö.; Abdulkadir, A.; Lienkamp, S. S.; Brox, T.; and Ronneberger, O. 2016. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 424–432. Springer.

Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; and Lu, H. 2019. Dual Attention Network for Scene Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3146–3154.

Funke, J.; Saalfeld, S.; Bock, D.; Turaga, S.; and Perlman, E. 2016. Miccai challenge on circuit reconstruction from electron microscopy images. *MICCAI. Springer*.

Funke, J.; Tschopp, F.; Grisaitis, W.; Sheridan, A.; Singh, C.; Saalfeld, S.; and Turaga, S. C. 2018. Large scale image segmentation with structured loss based deep learning for connectome reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7): 1669–1680.

Hatamizadeh, A.; Nath, V.; Tang, Y.; Yang, D.; Roth, H. R.; and Xu, D. 2021. Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images. In *MICCAI Brainlesion Workshop*, 272–284.

Hatamizadeh, A.; Tang, Y.; Nath, V.; Yang, D.; Myronenko, A.; Landman, B.; Roth, H. R.; and Xu, D. 2022. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 574–584.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-Excitation Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7132–7141.

Huang, W.; Deng, S.; Chen, C.; Fu, X.; and Xiong, Z. 2022. Learning to model pixel-embedded affinity for homogeneous instance segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 1007–1015.

Januszewski, M.; Kornfeld, J.; Li, P. H.; Pope, A.; Blakely, T.; Lindsey, L.; Maitin-Shepard, J.; Tyka, M.; Denk, W.; and Jain, V. 2018. High-precision automated reconstruction of neurons with flood-filling networks. *Nature Methods*, 15(8): 605–610.

Kasthuri, N.; Hayworth, K. J.; Berger, D. R.; Schalek, R. L.; Conchello, J. A.; Knowles-Barley, S.; Lee, D.; Vázquez-Reina, A.; Kaynig, V.; Jones, T. R.; et al. 2015. Saturated reconstruction of a volume of neocortex. *Cell*, 162(3): 648–661.

Kingma, D. P.; and Ba, J. 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Lee, K.; Zung, J.; Li, P.; Jain, V.; and Seung, H. S. 2017. Superhuman Accuracy on the SNEMI3D Connectomics Challenge. arXiv:1706.00120.

Li, J.; Wang, Y.; Huang, Y.; Wang, Z.; Zhang, Y.; Zhang, X.; Yang, Y.; Zhang, S.; and Yang, Y. 2024. VL-SAM: Vision-Language Segment Anything Model. In *Advances in Neural Information Processing Systems*, volume 37, 21877–21891.

Lin, Z.; Wei, D.; Lichtman, J.; and Pfister, H. 2021. PyTorch Connectomics: A Scalable and Flexible Segmentation Framework for EM Connectomics. arXiv:2112.05754.

Liu, X.; Cai, M.; Chen, Y.; Zhang, Y.; Shi, T.; Zhang, R.; Chen, X.; and Xiong, Z. 2024. Cross-dimension affinity distillation for 3d em neuron segmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11104–11113. IEEE Computer Society.

Luo, N.; Sun, R.; Pan, Y.; Zhang, T.; and Wu, F. 2024. Electron microscopy images as set of fragments for mitochondrial segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 3981–3989.

Ma, J.; He, Y.; Li, F.; Han, L.; You, C.; and Wang, B. 2024. Segment anything in medical images. *Nature Communications*, 15(1): 654.

Meilă, M. 2003. Comparing clusterings by the variation of information. In *Learning Theory and Kernel Machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003. Proceedings*, 173–187. Springer.

Rand, W. M. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336): 846–850.

Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; Mintun, E.; Pan, J.; Alwala, K. V.; Carion, N.; Wu, C.-Y.; Girshick, R.; Dollar, P.; and Feichtenhofer, C. 2025. SAM 2: Segment Anything in Images and Videos. In *The Thirteenth International Conference on Learning Representations*.

Roy, S.; Wald, T.; Koehler, G.; Rokuss, M. R.; Disch, N.; Holzschuh, J.; Zimmerer, D.; and Maier-Hein, K. H. 2023. SAM.MD: Zero-shot medical image segmentation capabilities of the Segment Anything Model. arXiv:2304.05396.

Sheridan, A.; Nguyen, T. M.; Deb, D.; Lee, W.-C. A.; Saalfeld, S.; Turaga, S. C.; Manor, U.; and Funke, J. 2023. Local shape descriptors for neuron segmentation. *Nature Methods*, 20(2): 295–303.

Sun, R.; Luo, N.; Pan, Y.; Mai, H.; Zhang, T.; Xiong, Z.; and Wu, F. 2023. Appearance Prompt Vision Transformer for Connectome Reconstruction. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 1423–1431.

Tang, C.; Chen, H.; Li, X.; Li, J.; Zhang, Z.; and Hu, X. 2021. Look closer to segment better: Boundary patch refinement for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13926–13935.

Wang, H.; Guo, S.; Ye, J.; Deng, Z.; Cheng, J.; Li, T.; Chen, J.; Su, Y.; Huang, Z.; Shen, Y.; et al. 2024. Sam-med3d: towards general-purpose segmentation models for volumetric medical images. In *Proceedings of the European Conference on Computer Vision*, 51–67. Springer.

Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; and Hu, Q. 2020. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11534–11542.

Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2018. Non-local Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7794–7803.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. CBAM: Convolutional Block Attention Module. In *Proceedings of the European Conference on Computer Vision*, 3–19.

Wu, J.; Wang, Z.; Hong, M.; Ji, W.; Fu, H.; Xu, Y.; Xu, M.; and Jin, Y. 2025. Medical sam adapter: Adapting segment anything model for medical image segmentation. *Medical Image Analysis*, 102: 103547.

Xu, C.; Wang, L.; Zhang, X.; and Sun, J. 2019. AC-FPN: Attention-Context Feature Pyramid Network for Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7547–7556.

Yang, L.; Zhang, Y.; Zhang, S.; Wang, H.; and Wang, J. 2019. DEA-Net: Detail-Enhanced Attention Network for Aerial Image Segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 6049–6058.

Yu, C.; Liu, T.; Li, A.; Qu, X.; Wu, C.; Liu, L.; and Hu, X. 2025. SAM-REF: Introducing Image-Prompt Synergy during Interaction for Detail Enhancement in the Segment Anything Model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 19356–19365.

Zhang, Y.; Guo, J.; Zhai, H.; Liu, J.; and Han, H. 2024a. SegNeuron: 3D Neuron Instance Segmentation in Any EM Volume with a Generalist Model. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 589–600. Springer.

Zhang, Y.; Yang, J.; Liu, Y.; Cheng, Y.; and Qi, Y. 2024b. Semisam: Enhancing semi-supervised medical image segmentation via sam-assisted consistency regularization. In *2024 IEEE International Conference on Bioinformatics and Biomedicine*, 3982–3986. IEEE.