Electron Microscopy Images as Set of Fragments for Mitochondrial Segmentation

Naisong Luo[1*], Rui Sun[1*], Yuwen Pan[1*], Tianzhu Zhang[1, 2†], Feng Wu[1, 2]

[1]Deep Space Exploration Laboratory/School of Information Science and Technology, University of Science and Technology of China
[2]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
{lns6, issunrui, panyw}@mail.ustc.edu.cn, {tzzhang, fengwu}@ustc.edu.cn

Abstract

Automatic mitochondrial segmentation enjoys great popularity with the development of deep learning. However, the coarse prediction raised by the presence of regular 3D grids in previous methods regardless of 3D CNN or the vision transformers suggest a possibly sub-optimal feature arrangement. To mitigate this limitation, we attempt to interpret the 3D EM image stacks as a set of interrelated 3D fragments for a better solution. However, it is non-trivial to model the 3D fragments without introducing excessive computational overhead. In this paper, we design a coherent fragment vision transformer (FragViT) combined with affinity learning to manipulate features on 3D fragments yet explore mutual relationships to model fragment-wise context, enjoying locality prior without sacrificing global reception. The proposed FragViT includes a fragment encoder and a hierarchical fragment aggregation module. The fragment encoder is equipped with affinity heads to transform the tokens into fragments with homogeneous semantics, and the multi-layer self-attention is used to explicitly learn inter-fragment relations with long-range dependencies. The hierarchical fragment aggregation module is responsible for hierarchically aggregating fragment-wise prediction back to the final voxel-wise prediction in a progressive manner. Extensive experimental results on the challenging MitoEM, Lucchi, and AC3/AC4 benchmarks demonstrate the effectiveness of the proposed method.

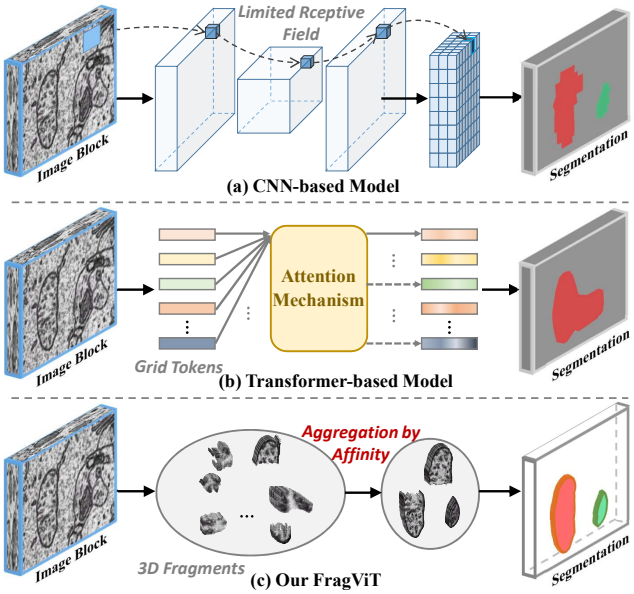

Figure 1: Compared to regular 3D-grid features or tokens of the CNN-based and transformer-based models, our FragViT model irregularly shaped fragments which contain more homogeneous semantics, resulting in finer segmentation.

Introduction

Mitochondria, as one of the crucial organelles, are the primary energy providers for cell activities and are indispensable for metabolism. Quantification of mitochondrial morphology is a fundamental task which can not only promote basic scientific research (e.g., electrophysiology (Ascoli 2002), cellular physiology (Donohue and Ascoli 2011)) but also provide new insight for clinical diagnosis (e.g., diabetes and neurodegenerative diseases (Martin 2012)). Electron microscopy (EM) is the only available imaging instrument with sufficient resolution to reveal mitochondrial detailed 3D geometry and perform dense segmentation without ambiguity. However, at this resolution, even moderately small 3D EM image stacks yield numerous mitochondrial numbers (e.g., typically hundreds to thousands of mitochondrial instances in a single megapixel image (Wei et al. 2020)) that are prohibitively laborious for human manual annotation. Recently, with the development of deep learning, considerable works (Wei et al. 2020; Nightingale et al. 2021; Li et al. 2021a) have turned their attention to deep neural networks in pursuit of automatic mitochondria segmentation. Since all mitochondrial instances are of the same type (i.e., biological organelle), with intricate morphology and densely intertwined branches, how to fully probe discriminative information to perform accurate segmentation is thus extremely challenging.

To tackle the mitochondria segmentation problem, existing methods can be roughly categorized as 3D convolutional neural networks (CNN) and the vision transformer (ViT) paradigms. On the one hand, 3D CNN paradigms such as 3D U-Net (Wei et al. 2020; Li et al. 2021a) dominate

*These authors contributed equally.
†Corresponding author.

this field credited to their simplicity yet competitive performance. However, 3D CNN is not the out-of-the-box solution for mitochondrial segmentation considering the following natures: (1) Local context. The inherent locality of 3D CNN compromises its capability to model long-range dependencies and capture global voxel context, leading to suboptimal results. (2) Coarse prediction. Considering that 3D CNN manipulates features on regular grids (Cartesian layout) in an inflexible manner, it doesn't adhere to the semantic structure of the physical world (Fig. 1 (a)). Additionally, typical CNN visual backbones have large strides, which inevitably carry cluttered semantics within grid cells. To make matters worse, the negative impact is inevitably amplified by the ambiguity in grid cells raised by the numerous boundaries between mitochondrial instances. On the other hand, ViT (Hatamizadeh et al. 2022; Tang et al. 2022) has gained increasing attention due to its long-range modeling ability derived from the multi-layer self-attention against the local context of CNN (Fig. 1 (b)). However, ViT extracts 2D coarse features (tokens) exactly as its convolutional counterpart does via patch partition, lured into the trap of coarse predictions. Then, the question naturally arises: The regular 3D grids may not be the optimal feature arrangement, and how to discover better solutions to make each cell contain more homogeneous semantics?

After the in-depth analysis, we attempt to interpret the 3D EM image stacks as a set of interrelated 3D *fragments*, where the fragments indicate a group of adjacent voxels with homogeneous semantics. We argue that 3D fragments matter in mitochondria segmentation, which is intuitively sensible from the definition of the task itself; that is, 3D fragments have flexible geometrics and carry compact semantics compared to regular grid cells, adapting to the considerable boundaries that exist between mitochondrial instances. However, it is non-trivial to model the 3D fragments without introducing excessive computational overhead. In this paper, we analyze the coarse prediction problem of previous methods, regardless of 3D CNN or ViT paradigms, and shed light on the possibility of closer collaboration between modeling 3D fragments and the ViT (Fig. 1 (c)). Specifically, we design a coherent **Fragment Vision Transformer (FragViT)** to *partition electron microscopy images into the fragment set for robust mitochondria segmentation*. To model the 3D fragments with light computation costs, We draw inspiration from affinity learning (Huang et al. 2022) to seek perceptibility to spatial position and discrimination to surrounding voxels, that is, implicitly describing fragment geometrics and ensuring the tessellation of the entire fragment set enables us to conduct semantic segmentation by per-fragment prediction.

Our FragViT consists of a fragment encoder and a hierarchical fragment aggregation module. The **fragment encoder** is equipped with affinity heads which distribute each voxel to 27 nearest neighbor tokens. With the help of the learned affinity, the region corresponding to each token can aggregate voxels with homogeneous semantics to form a fragment, which can be regarded as a "superpixel". And the multi-layer self-attention is used to explicitly learn inter-fragment relations with long-range dependencies. The **hierarchical fragment aggregation module** is responsible for hierarchically aggregating fragment-wise prediction based on the encoded fragment embedding back to the final voxel-wise prediction in a progressive manner. In this way, FragViT manipulates features on 3D fragments yet explores mutual relationships to model fragment-wise context, enjoying locality prior without sacrificing global reception.

The contributions of our method can be summarized as follows: (1) We analyze the coarse prediction problem of previous methods, both 3D CNN and vision transformers paradigms, and shed light on the possibility of closer collaboration between modeling 3D fragments and the vision transformers. (2) We propose a novel Fragment Vision Transformer (FragViT) combined with affinity learning to manipulate features on 3D fragments yet explore mutual relationships to model fragment-wise context, enjoying locality prior without sacrificing global reception. (3) Extensive experimental results on the challenging MitoEM, Lucchi, and AC3/AC4 benchmarks demonstrate the effectiveness of the proposed method.

Related Work

Mitochondria Segmentation

Mitochondria segmentation is important for the biological study of cellular functions and subcellular activities. Early works are mainly based on traditional image processing techniques (Vazquez-Reina et al. 2011; Lucchi et al. 2014; Seyedhosseini, Ellisman, and Tasdizen 2013). With significant advances in deep learning (DL), recent DL-based autosegmentation methods (Mai et al. 2023; Wangkai et al. 2023; Pan et al. 2023; Sun et al. 2023b) demonstrates huge performance gains, which can be roughly categorized as convolutional neural network (CNN) and the vision transformer (ViT) paradigms. For example, U3D-BC (Wei et al. 2020) utilizes a supervised 3D U-Net (Çiçek et al. 2016) architecture to predict foreground and contour followed by a post-processing step to produce final instance segmentation. After that, ResUNet (Li et al. 2021a) proposes the asymmetric 3D convolution blocks to adapt the anisotropy of EM images. To alleviate the limited receptive fields of CNN, recent ViT-based methods (Xie et al. 2023; Hatamizadeh et al. 2022) utilize a transformer as the encoder to learn sequence representations of the input volume and effectively capture the global multi-scale information, achieving improved performance. However, regardless of 3D CNN or ViT, their regular 3D-grid features inevitably result in coarse predictions. Different from these methods, we propose the concept of irregular 3D fragment, which is more compatible with the nature of EM images.

Vision Transformer

Transformer (Vaswani et al. 2017) is first introduced in the field of nature language processing (Kitaev, Kaiser, and Levskaya 2020; Yang et al. 2019) for machine translation. Since its remarkable success, ViT (Dosovitskiy et al. 2020) applies the transformer to vision tasks as a generic vision backbone. Specifically, an image can be divided by the grid into patches of pixels, and each patch can be flattened and

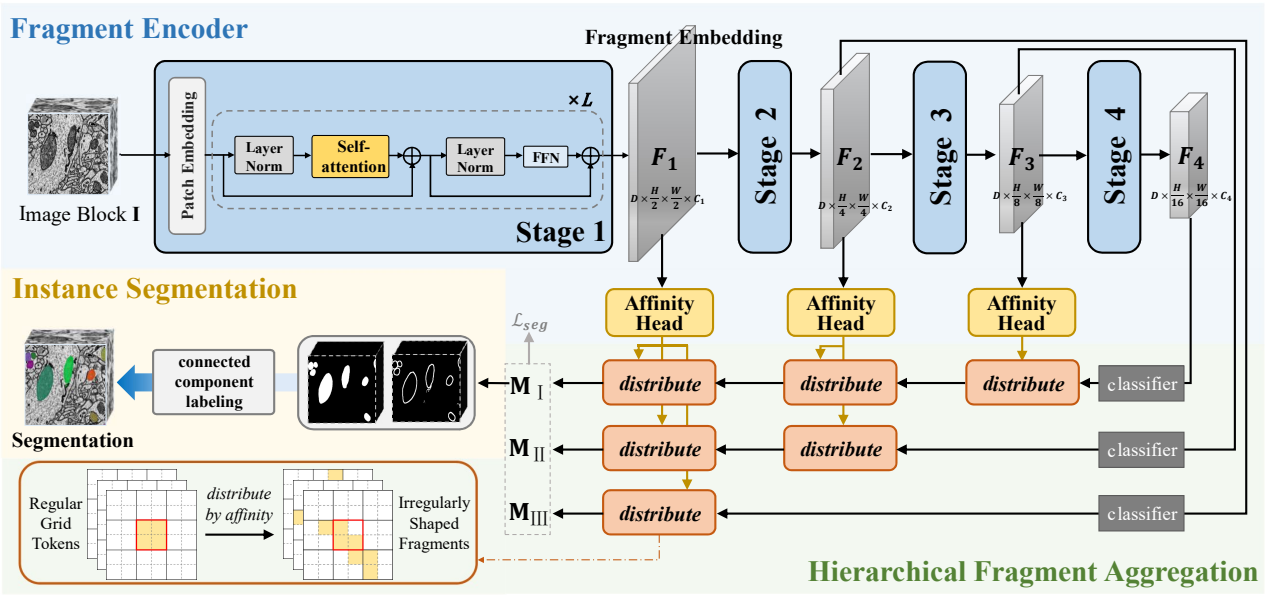

Figure 2: The framework of our proposed FragViT. The fragment encoder is responsible for producing the pyramid tokens which serve as fragment embedding in a voxel-to-token association manner by learned affinity. The hierarchical fragment aggregation aggregates the fragments to generate semantic masks and instance boundaries. Finally, an instance segmentation module generates mitochondria segmentation results using connected component labeling.

projected into a vector. This vector is a token that encodes the visual information of that patch. The ViT uses these tokens as inputs and learns to capture the global and local relationships among them by the self-attention operation. MaskFormer (Cheng, Schwing, and Kirillov 2021) and Mask2Former (Cheng et al. 2022) apply the global interaction of the transformer to the segmentation task, opening up a new segmentation paradigm (Wang et al. 2022; Luo et al. 2023; Wang, Sun, and Zhang 2023; Sun et al. 2022, 2023d; Wang, Luo, and Zhang 2023; Sun et al. 2023c). UNETR (Hatamizadeh et al. 2022) and Swin UNETR (Tang et al. 2022) demonstrate that transformers can also take the lead in biological image segmentation tasks (Sun et al. 2021, 2023a). Our modeled fragment borrows the form of the token in the transformer, and the affinity learned at the shallow level acts on the deeper tokens and can in turn benefit the transformer by making the feature representation more semantically consistent.

Method

Overview

FragViT mainly consists of a fragment encoder and a hierarchical fragment aggregation module. The fragment encoder, which is transformer-based, is responsible for producing the pyramid tokens that serve as fragment embedding. Based on the learned fragment embedding from each stage, the hierarchical fragment aggregation module is responsible for aggregating the fragments to generate semantic masks and instance boundaries. Finally, an instance segmentation module generates mitochondria segmentation results using connected component labeling.

Transformer as Fragment Encoder

Current state-of-the-art mitochondria segmentation models (Wei et al. 2020; Li et al. 2021a) use 3D CNN (Ji et al. 2012) to extract multi-scale features. Compared to CNN, the vision transformer has stronger long-range modeling capability since the self-attention mechanism in the transformer can aggregate contexts with a global receptive field. In this work, we take full advantage of the transformer architecture to learn inter-region relations using the self-attention mechanism to learn fragment embedding. In specific, the input volume $I \in \mathbb{R}^{D\times H\times W}$ is cropped from electron microscopy (EM) images, where D, H and W denote the depth, height, and width, respectively. In the first stage of the encoder, we first divide I into $D\times\frac{H}{2}\times\frac{W}{2}$ patches, each of size $1\times2\times2$. Considering the anisotropy of the electron microscopy images, i.e., the resolution in the depth direction (D) is significantly smaller than the other two directions (H, W), we will not reduce the resolution in the depth direction when dividing the patches, in order to avoid corrupting the context of the depth direction. Then, we feed the flattened patches to a linear projection and obtain embedded patches (i.e., tokens) of size $\frac{DHW}{2^2}\times C_1$. After that, the tokens along with a 3D position embedding are passed through an encoder block, and the output restores the spatial resolution and yields a feature map $\mathbf{F}_1$ of size $D\times\frac{H}{2}\times\frac{W}{2}\times C_1$. In the same way, at the beginning of each stage i, using the feature map from the previous stage as input, we can obtain multi-scale feature maps of $\mathbf{F}_i$ of size $D\times\frac{H}{P^i}\times\frac{W}{P^i}\times C_i$, where $P^i=2^i$, and $i=\{1,2,3,4\}$.

Given the embedded feature sequence $\mathbf{F}_i\in\mathbb{R}^{\frac{DHW}{(P^i)^2}\times C_i}$ in the stage i, we can get the tokens by mapped queries $\mathbf{Q}_i$,

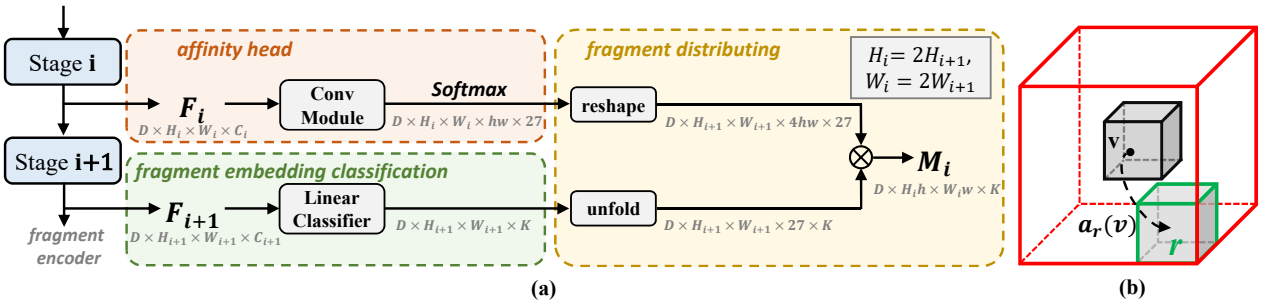

Figure 3: Illustration of model details. (a) The fragment aggregation process, including the affinity head, fragment embedding classification, and fragment distributing. (b) Affinity $a_{\boldsymbol{r}}(\boldsymbol{v})$ between voxel $\boldsymbol{v}$ and neighbor token region $\boldsymbol{r}$ in 3D volume.

keys $\mathbf{K}_i$ and values $\mathbf{V}_i$ by the self-attention mechanism, as:

$$\text{SelfAtt}(\mathbf{F}_i) = \text{Softmax}\left(\frac{\mathbf{Q}_i\mathbf{K}_i^\top}{\sqrt{C}}\right)\mathbf{V}_i, \tag{1}$$

where $\sqrt{C}$ is a scaling factor and $\top$ denotes the transpose operation. The Eq. 1 is implemented with the multi-head mechanism and we apply the feed-forward network (FFN) to obtain the final output following the standard transformer (Vaswani et al. 2017). The tokens output by the transformer have the capability of long-range correlation, but we expect them to be aggregated into the desired fragment embedding, rather than regular grid features. Next, we describe why tokens can be regarded as fragment embeddings.

Learning Fragment Embedding

The regions corresponding to the token sequences of the transformer are regular grids. Grid-like regions often contain both confusing mitochondrial foreground and background, which can lead to semantic ambiguity in the token. Inspired by related affinity learning methods (Huang et al. 2022; Ru et al. 2022), we propose to use local affinity to model voxel-to-token connection

Specifically, the feature map $\mathbf{F}_i \in \mathbb{R}^{D\times H_i\times W_i\times C_i}$ (or $N_i = D \times H_i \times W_i$ tokens) output from stage i of the fragment encoder can correspond to a $D\times H_i\times W_i$ 3D grid. Each token lays on a single cell which serves as a "placeholder" of its corresponding fragment $\boldsymbol{r}$. Note that the cell itself is just a token position indicator and has nothing to do with the shape of the actual fragment. The voxel-to-token connection is built by assigning each voxel $\boldsymbol{v}$ located at (z, y, x) to fragment $\boldsymbol{r}$ with a probability $a_{\boldsymbol{r}}(\boldsymbol{v})$. We make voxel $\boldsymbol{v}$ be associated only with surrounding tokens, which satisfies:

$$\sum_{\boldsymbol{r}\in\mathcal{N}_{\boldsymbol{v}}} a_{\boldsymbol{r}}(\boldsymbol{v}) = 1, \tag{2}$$

where $\mathcal{N}_{\boldsymbol{v}}$ is the neighborhood of voxel $\boldsymbol{v}$, and we set its size to $3 \times 3 \times 3$, thus $|\mathcal{N}_{\boldsymbol{v}}| = 27$. As illustrated in Figure 3(b), the voxel $\boldsymbol{v}$ is assigned to the placeholder of fragment $\boldsymbol{r}$ (greed box) with probability $a_{\boldsymbol{r}}(\boldsymbol{v})$ within its $3\times3\times3$ neighborhood (red box). Conversely, the shape of the fragment $\boldsymbol{r}$ can be defined as the probability distribution of all voxels in its 27 neighborhood. We use an affinity matrix $\mathbf{A}^i \in \mathbb{R}^{D\times H_i\times W_i\times(hw)\times 27}$ to formulate all fragments corresponding to $N_i = D \times H_i \times W_i$ tokens, where hw refers to the number of voxels contained in a patch. To predict the affinity matrix $\mathbf{A}$, we design a lightweight affinity head which consists of one $3 \times 3 \times 3$ 3D convolution layer and a $1 \times 1 \times 1$ convolution layer followed by a Softmax operator to produce normalized probabilities as in Eq. 2, as depicted in Figure 3(a). The affinity stands for the voxel-to-token association, which is used to aggregate voxels into fragments.

In order to describe the mechanism of the fragment more clearly, we make a visualization in the bottom left of Figure 2. Given a set of tokens (solid line) where we assume each token contains 2×2 voxels (dashed line), we learn the affinity matrix to distribute each voxel to 27 nearest neighbor tokens. In this way, the region corresponding to each token can aggregate voxels with homogeneous semantics to form a fragment, which can be regarded as a "superpixel". The yellow area is the fragment formed by the middle token (red box).

Hierarchical Fragment Aggregation

Most previous works (Wei et al. 2020; Li et al. 2021a) use UNet-style (Ronneberger, Fischer, and Brox 2015) encoder-decoder structures to predict segmentation mask. Instead, with the fragment embedding, we can directly employ the clustering of learned fragments for segmentation. In specific, we apply a linear classifier on fragment embedding $\mathbf{F}_i$, producing class logits $\mathbf{Y}_i \in \mathbb{R}^{N_i\times K}$ for all tokens, where K is number of classes. Then, we can assign voxel $\boldsymbol{v}$ to the tokens located in $\mathcal{N}_{\boldsymbol{v}}$ based on the affinity $\mathbf{A}$. After that, we *distribute* the token logits corresponding voxels by weight to get voxel-level classification results. For the voxel $\boldsymbol{v} = (z, y, x)$ from previous stage $j (j < i)$, the class logits is calculated by:

$$\text{S}[j] \xleftarrow{(h,w)} \text{S}[i] : \mathbf{M}_j[z, y, x] = \sum_{\boldsymbol{r}\in\mathcal{N}_{\boldsymbol{v}}} \mathbf{Y}_i(\boldsymbol{r}) \cdot a^j_{\boldsymbol{r}}(\boldsymbol{v}), \tag{3}$$

where S denotes stage, $\mathbf{Y}_i(\boldsymbol{r})$ is the class logits of the fragment embedding corresponding to fragment $\boldsymbol{r}$, $a^j_{\boldsymbol{r}}(\boldsymbol{v})$ is the affinity (i.e., assignment probability) from $\mathbf{A}_j$, and $\mathbf{M}_j \in \mathbb{R}^{D\times(H_jh)\times(W_jw)\times K}$ is the output logits map with higher resolution compared to the stage i. (h, w) on the arrow implies the increased height, width resolution relative

to stage j in the aggregation process. Figure 3(a) exhibits an example of the fragment embedding classification and following $\mathrm{S}[i] \overset{(h,w)}{\longleftarrow} \mathrm{S}[i+1]$ distributing in detail. This process can be regarded as fine-grained voxels aggregating into coarse-grained fragment embeddings from an opposite direction. Experiments show that the one-step fragment aggregation results of $\mathbf{F}_4$ (i.e., $\mathrm{S}[1] \overset{(2,2)}{\longleftarrow} \mathrm{S}[4]$) performs poorly, so we aggregate volume regions stage-by-stage based on hierarchical features as the segmentation results:

$$\mathbf{M}_{\mathrm{I}} = \mathrm{S}[1] \overset{(2,2)}{\longleftarrow} \mathrm{S}[2] \overset{(1,1)}{\longleftarrow} \mathrm{S}[3] \overset{(1,1)}{\longleftarrow} \mathrm{S}[4], \tag{4}$$

where $\mathbf{M}_{\mathrm{I}} \in \mathbb{R}^{D \times H \times W \times K}$ can serve as the voxel logits. As shown in Figure 2, we use two other paths to obtain segmentation results $\mathbf{M}_{\mathrm{II}}, \mathbf{M}_{\mathrm{III}}$ for assisting training of fragment embedding learners.

Training and Inference

Training Objectives. We use a linear combination of a binary cross-entropy loss and a dice loss (Milletari, Navab, and Ahmadi 2016) to constrain the prediction:

$$\begin{aligned} \mathcal{L}_{seg}(\mathbf{M}_{\mathrm{I}}) &= \lambda_1 \mathcal{L}_{bce}(\mathbf{M}_{\mathrm{I}}, \hat{\mathbf{M}}) + \lambda_2 \mathcal{L}_{dice}(\mathbf{M}_{\mathrm{I}}, \hat{\mathbf{M}}), \\ \mathcal{L}_{total} &= \mathcal{L}_{seg}(\mathbf{M}_{\mathrm{I}}) + \mathcal{L}_{seg}(\mathbf{M}_{\mathrm{II}}) + \mathcal{L}_{seg}(\mathbf{M}_{\mathrm{III}}), \end{aligned} \tag{5}$$

where $\hat{\mathbf{M}}$ is the ground-truth and λ_1, λ_2 are coefficients.

Instance Segmentation Inference. The goal of the mitochondrial segmentation task is to obtain each mitochondrial instance. The network predicts semantic masks of the foreground as well as the boundary of the targets, i.e., $K = 2$, followed by a heuristic instance segmentation module to split individual instances. Concretely, the boundary mask is subtracted from the foreground mask to get the seed map of the instance. The foreground mask and seed map are then fed into the seeded watershed algorithm (Lin et al. 2021) to obtain the instance segmentation mask.

Experiments

Dataset

To demonstrate the effectiveness of our proposed model, we conduct extensive experiments on three 3D EM image benchmarks: MitoEM (Wei et al. 2020), Lucchi (Lucchi et al. 2011) and AC3/AC4 (Arganda-Carreras, Turaga, and Berger 2015). These three datasets belong to different tasks (e.g., mitochondria instance/semantic segmentation and neuron instance segmentation), and the different tasks have their own baselines and metrics. The division of these datasets follows previous work. Due to the limited labeled EM data, such a data division is widely used to evaluate model performance and we did not modify it for fair comparison. We complete the different tasks uniformly with a single model (FragViT) and compare them with the task-specific approaches using respective metrics.

MitoEM is a mitochondria instance dataset. It is divided into two subsets, MitoEM-R (rat tissue) and MitoEM-H (human tissue), whose resolution is anisotropic $30 \times 8 \times 8$ nm. The size of the training set is $400 \times 4096 \times 4096$ and the validation set is $100 \times 4096 \times 4096$ voxels. We adopt AP, AP_{50} and AP_{75} scores as evaluation metrics to quantify the effectiveness of instance segmentation. We also show the results on AP_s, AP_m and AP_l metrics, which denote AP_{75} performance of small, medium, and large mitochondria instances (divided by 5K and 15K voxels), respectively.

Lucchi is a mitochondria semantic segmentation dataset. The training and testing data volumes are both $165 \times 1024 \times 768$ with isotropic $5 \times 5 \times 5$ nm resolution. We adopt the metric of Jaccard-index coefficient (Jaccard) and dice similarity coefficient (DSC) to evaluate the effectiveness of semantic segmentation.

AC3/AC4 consists of EM images of mouse somatosensory cortex neuron instances, where the size of AC3 is $256 \times 1024 \times 1024$ and AC4 consists of $100 \times 1024 \times 1024$ voxels with $29 \times 3 \times 3$ nm resolution. Following the SNEMI3D challenge, we use the top 80 slices of AC4 as the training set and the rest of AC4 as the validation set. The top 100 slices of AC3 are testing set. Following the conventions (Funke et al. 2018), we report VOI (variation of information) and ARAND (adapted Rand error) to evaluate the results. And VOI_s and VOI_m represent the split and merge error of predicted instances, respectively. Smaller values of the above metrics indicate better segmentation performance.

Implementation Details

In the fragment encoder, the number of layers is {2, 2, 2, 2} in each stage. The size of the input volume is anisotropic (18, 160, 160). We set $\lambda_1 = 1$ and $\lambda_2 = 0.5$. During training, our model is trained with a batch size of 2, using the Adam optimizer with an initial learning rate of 0.0001 on two NVIDIA RTX 3090 GPUs for 160,000 iterations.

Main Results

Results on MitoEM. Tab 1 reports the 3D mitochondria segmentation performance of our method and current segmentation methods on MitoEM-R and MitoEM-H datasets. There are U2D-B, U3D-A, U3D-BC (Wei et al. 2020), Nightingale (Nightingale et al. 2021), CLMS (Li et al. 2021b), ResUNet (Li et al. 2021a) with 3D CNN backbone, and MaskFormer (Cheng, Schwing, and Kirillov 2021), Mask2Former (Cheng et al. 2022), UNETR (Hatamizadeh et al. 2022), Swin UNETR (Tang et al. 2022) with vision transformer backbone. The proposed FragViT consistently outperforms the state-of-the-art methods by a large margin on both datasets. Compared to the CNN-based methods, our approach outperforms the previous best method ResUNet by **4.1%** on MitoEM-R and **6.4%** on MitoEM-H in AP metric, indicating that features with long-range dependencies captured by the transformer are more favorable for EM volume segmentation compared to CNN features. And our approach surpasses all of the transformer-based methods, demonstrating that the fragment we learned is able to cope with irregularly shaped mitochondria. Besides, we present the qualitative results of mitochondria instance segmentation and corresponding 3D reconstruction in Figure 4, demonstrating the effectiveness of FragViT.

Results on Lucchi. Since the Lucchi dataset is isotropic, we use a $2 \times 2 \times 2$ patch size in each stage. And the network

Method	Backbone	MitiEM-R						MitoEM-H					
		AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
U2D-B	CNN	28.4	40.2	35.5	10.4	62.8	48.1	36.8	62.3	59.7	40.8	81.4	71.1
U3D-A		26.5	38.4	32.8	40.8	23.5	65.3	42.1	65.5	61.7	56.4	77.4	61.7
U3D-BC		45.6	57.3	52.1	29.0	75.1	49.0	45.5	66.2	60.5	48.9	82.0	61.8
Nightingale		-	-	71.5	0.70	40.4	78.7	-	-	62.5	3.40	47.8	73.4
CLMS		-	89.5	87.0	20.3	74.3	91.3	-	82.8	78.7	29.6	77.8	83.0
ResUNet†		72.8	90.8	89.7	26.0	83.9	92.8	61.1	82.4	79.9	38.4	83.3	83.9
MaskFormer†	Transformer	60.4	88.8	75.2	10.1	66.9	79.7	55.3	80.6	69.0	21.4	72.7	74.5
Mask2Former†		64.5	91.0	18.9	11.8	71.3	82.7	79.5	83.1	72.9	23.5	75.3	72.2
UNETR†		71.7	91.4	86.0	19.8	83.7	89.5	60.2	85.8	76.8	23.9	76.4	79.0
Swin UNETR†		73.1	92.7	90.4	32.9	85.1	92.4	61.5	87.6	80.3	37.9	82.2	84.2
Ours		**76.9**	**95.8**	**92.3**	**39.4**	**87.8**	**94.2**	**67.5**	**90.7**	**82.9**	**42.3**	**84.5**	**85.6**

Table 1: Comparisons of existing mitochondria instance segmentation methods and some general segmentation approaches on MitoEM-R and MitoEM-H validation set. † denotes the performance of our reproduction by adopting the official code.

Method	Jaccard	DSC
Lucchi	75.5	86.0
Liu	86.4	92.6
Yuan	86.5	92.7
Wei	88.7	-
Casser	89.0	94.2
ResUNet	89.5	94.5
Ours	**90.5**	**95.0**

Table 2: Mitochondria semantic segmentation on Lucchi.

Method	VOI	VOI_s	VOI_m	ARAND
SuperHuman	1.433	1.091	0.342	0.169
MALA	1.343	1.099	0.245	0.109
PEA	1.205	0.912	0.293	0.121
Ours	**1.054**	**0.868**	**0.191**	**0.093**

Table 3: Neuron instance segmentation results on AC3/AC4. Note that smaller values indicate better performance.

predicts the foreground of the mitochondria only for semantic segmentation, i.e., $K = 1$. As shown in Table 2, among the previous methods such as Lucchi (Lucchi et al. 2011), Liu (Liu et al. 2020), Yuan (Yuan et al. 2020), Wei (Wei et al. 2020), Casser (Casser et al. 2020), and ResUNet (Li et al. 2021a), our method achieves a consistent performance gain on both metrics attributed to fragment modeling, unlike the grid features that leads to coarse segmentation.

Results on AC3/AC4. Since the neurons in AC3/AC4 are in close proximity to each other and the foreground occupies almost the entire volume, we adjust the output classes of FragViT to the dense affinity in 3 directions for better instance discrimination, i.e., $K = 3$. As shown in Table 3, compared to previous neuron segmentation methods that belong to the UNet family, such as SuperHuman (Lee et al. 2017), MALA (Funke et al. 2018), and PEA (Huang et al. 2022), FragViT yields segmentation results with fewer split and merge errors, proving that our approach can effectively generalize to other volume segmentation tasks.

Configuration	AP	Params
baseline	73.0	29.9M
baseline+fragment	74.2	32.0M
baseline+fragment+hierarchical	76.9	34.2M

Table 4: Ablation of main components on MitoEM-R.

Ablation Study

We conduct comprehensive ablation studies with FragViT on the validation set of the MitoEM-R dataset. Note that we replace the hierarchical fragment aggregation module with a conventional UNet-style decoder and maintain the transformer-based encoder as our baseline.

Effectiveness of Main Components. Table 4 summarizes the results of module ablation studies under different configurations. (1) Compared to the baseline, fragment brings 1.2% performance improvement in AP. The improvements can be mainly ascribed to the strong ability of the fragment to portray the adaptive volume region, exploiting the potential of the fragment embedding to complement sequence-to-sequence embedding. (2) Then we investigate the impact of introducing hierarchical fragment aggregation and observe an absolute performance lift (from 74.2 to 76.9 in AP). The improvements suggest that the hierarchical design stimulates the potential of the learned affinity to aggregate complex mitochondria fragments. (3) The tremendous performance gains are accompanied by a negligible increase in parameters, proving the simplicity and effectiveness of our fragment modeling. Besides, we present parameters of our model, which achieves significant performance gains with little increase in computational overhead.

Scope of the neighborhood $\mathcal{N}_v$. We further investigate the impact of the scope of the neighborhood in Table 5. It can be observed that the $\mathcal{N}_v$ of $3\times3\times3$ achieves better performance than that of $1 \times 3 \times 3$ or 6-nearest, which demonstrated the effectiveness of our design.

Optimal Hierarchical Fragment Aggregation Path. As the effectiveness of the hierarchical fragment aggregation has been approved in Table 4, we then explore the optimal

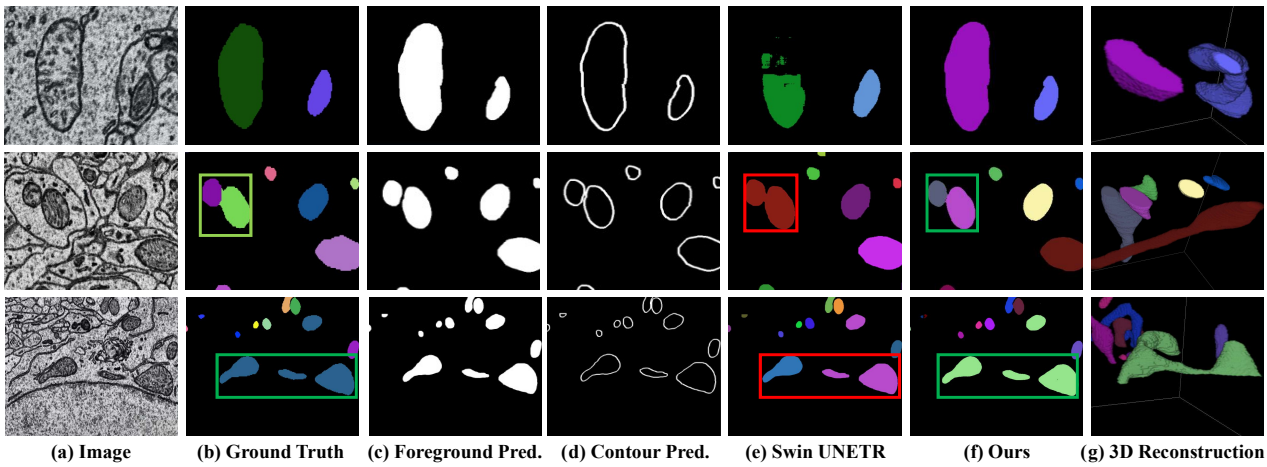

Figure 4: Qualitative results of different methods on the MitoEM-R validation set. It can be seen that, Swin UNETR appears to have a broken foreground, while FragViT can activate the intact target. When mitochondria crowd the space, our method can accurately distinguish between neighboring instances.

$\mathcal{N}_v$	AP
$1 \times 3 \times 3$	75.3
6-nearest	76.2
$3 \times 3 \times 3$ **(Ours)**	**76.9**

Table 5: Results of different neighbors.

aggregation path	AP
$\mathrm{S}[1] \xleftarrow{(2,2)} \mathrm{S}[4]$	74.2
$\mathrm{S}[1] \xleftarrow{(2,2)} \mathrm{S}[2] \xleftarrow{(1,1)} \mathrm{S}[3]$	74.7
$\mathrm{S}[2] \xleftarrow{(4,4)} \mathrm{S}[3] \xleftarrow{(1,1)} \mathrm{S}[4]$	75.4
$\mathrm{S}[1] \xleftarrow{(2,2)} \mathrm{S}[2] \xleftarrow{(1,1)} \mathrm{S}[3] \xleftarrow{(1,1)} \mathrm{S}[4]$	**76.9**

Table 6: Comparisons of different aggregation paths.

aggregation path. As shown in Table 6, when we distribute logits from stage 3, passing stage 2 to stage 1 (3 steps), the performance lifts because the fragment embedding aggregates more context from stage 2 and stage 3. Next, when we choose the path that starts from deeper stage 4, we gain more performance improvement. This is because deeper fragment embedding carries consistent semantics without confusion. When we go through the complete process of hierarchical aggregation from stage 4 to stage 1, we obtain the highest performance.

Explainable Visualization Study

We analyze the semantical consistency of fragments, which is described using "fragment entropy". Specifically, we count the category probabilities of the voxels involved in feature aggregation by 3D ViT and our FragViT. Then we calculate the per-voxel logit entropy, and plot the histogram within each grid or fragment as shown in the left part of Figure 5. Compared to grid, our constructed fragments are

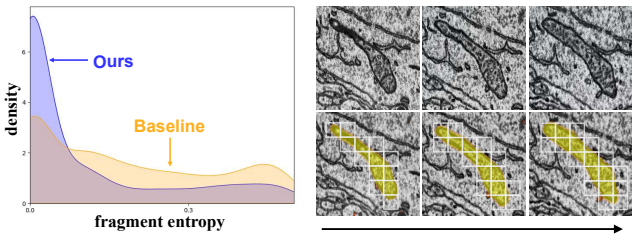

Figure 5: (Left) Distribution of fragment entropies. (Right) Visualizations of the fragments (yellow indicates high affinity) in 3 consecutive EM slices and their corresponding tokens (marked using white cell).

able to absorb more homogeneous semantic voxels, which is consistent with our motivation. In the right part of Figure 5, we visualize the fragment affinity (from the shallow stage) across 3 consecutive slices on the MitoEM-R validation set. We can observe that the learned affinity still captures fine-grained boundaries of mitochondria and accurately distinguishes the foreground from the confusing background. This suggests that our fragment embeddings carry more consistent semantics than traditional CNN features or general transformer tokens.

Conclusion

In this paper, we propose a novel network Fragment Vision Transformer (FragViT) to model the fragment (3D volume region) in a lightweight way for mitochondria segmentation in electron microscopy images, which is more compatible with the nature of EM images. Specifically, we design a coherent FragViT to learn featuresand explore mutual relationships to model fragment-wise context. The hierarchical fragment aggregation module is responsible for aggregating fragment-wise prediction in a hierarchical manner. Extensive experiments show the effectiveness of our method.

Acknowledgments

This work was partially supported by the the Youth Innovation Promotion Association CAS (Grant 2018166), National Nature Science Foundation of China (Grant 62021001, 62071122).

References

Arganda-Carreras, I.; Turaga, S. C.; and Berger. 2015. Crowdsourcing the creation of image segmentation algorithms for connectomics. *Frontiers in neuroanatomy*, 142.

Ascoli, G. A. 2002. *Computational neuroanatomy: Principles and methods*. Springer Science & Business Media.

Casser, V.; Kang, K.; Pfister, H.; and Haehn, D. 2020. Fast mitochondria detection for connectomics. In *Medical Imaging with Deep Learning*, 111–120. PMLR.

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.

Cheng, B.; Schwing, A.; and Kirillov, A. 2021. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34: 17864–17875.

Çiçek, Ö.; Abdulkadir, A.; Lienkamp, S. S.; Brox, T.; and Ronneberger, O. 2016. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, 424–432. Springer.

Donohue, D. E.; and Ascoli, G. A. 2011. Automated reconstruction of neuronal morphology: an overview. *Brain research reviews*, 67(1-2): 94–102.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

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.; 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.

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.

Ji, S.; Xu, W.; Yang, M.; and Yu, K. 2012. 3D convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1): 221–231.

Kitaev, N.; Kaiser, Ł.; and Levskaya, A. 2020. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*.

Lee, K.; Zung, J.; Li, P.; Jain, V.; and Seung, H. S. 2017. Superhuman accuracy on the SNEMI3D connectomics challenge. *arXiv preprint arXiv:1706.00120*.

Li, M.; Chen, C.; Liu, X.; Huang, W.; Zhang, Y.; and Xiong, Z. 2021a. Advanced Deep Networks for 3D Mitochondria Instance Segmentation. *arXiv preprint arXiv:2104.07961*.

Li, Z.; Chen, X.; Zhao, J.; and Xiong, Z. 2021b. Contrastive learning for mitochondria segmentation. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 3496–3500. IEEE.

Lin, Z.; Wei, D.; Petkova, M. D.; Wu, Y.; Ahmed, Z.; Zou, S.; Wendt, N.; Boulanger-Weill, J.; Wang, X.; Dhanyasi, N.; et al. 2021. NUCMM dataset: 3d neuronal nuclei instance segmentation at sub-cubic millimeter scale. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 164–174. Springer.

Liu, J.; Li, L.; Yang, Y.; Hong, B.; Chen, X.; Xie, Q.; and Han, H. 2020. Automatic reconstruction of mitochondria and endoplasmic reticulum in electron microscopy volumes by deep learning. *Frontiers in neuroscience*, 14: 599.

Lucchi, A.; Márquez-Neila, P.; Becker, C.; Li, Y.; Smith, K.; Knott, G.; and Fua, P. 2014. Learning structured models for segmentation of 2-D and 3-D imagery. *IEEE transactions on medical imaging*, 34(5): 1096–1110.

Lucchi, A.; Smith, K.; Achanta, R.; Knott, G.; and Fua, P. 2011. Supervoxel-based segmentation of mitochondria in em image stacks with learned shape features. *IEEE transactions on medical imaging*, 31(2): 474–486.

Luo, N.; Pan, Y.; Sun, R.; Zhang, T.; Xiong, Z.; and Wu, F. 2023. Camouflaged Instance Segmentation via Explicit De-Camouflaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17918–17927.

Mai, H.; Sun, R.; Zhang, T.; Xiong, Z.; and Wu, F. 2023. DualRel: Semi-Supervised Mitochondria Segmentation From a Prototype Perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19617–19626.

Martin, L. J. 2012. Biology of mitochondria in neurodegenerative diseases. *Progress in molecular biology and translational science*, 107: 355–415.

Milletari, F.; Navab, N.; and Ahmadi, S.-A. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, 565–571. Ieee.

Nightingale, L.; de Folter, J.; Spiers, H.; Strange, A.; Collinson, L. M.; and Jones, M. L. 2021. Automatic instance segmentation of mitochondria in electron microscopy data. *bioRxiv*.

Pan, Y.; Luo, N.; Sun, R.; Meng, M.; Zhang, T.; Xiong, Z.; and Zhang, Y. 2023. Adaptive template transformer for mitochondria segmentation in electron microscopy images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 21474–21484.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 234–241. Springer.

Ru, L.; Zhan, Y.; Yu, B.; and Du, B. 2022. Learning affinity from attention: End-to-end weakly-supervised semantic segmentation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16846–16855.

Seyedhosseini, M.; Ellisman, M. H.; and Tasdizen, T. 2013. Segmentation of mitochondria in electron microscopy images using algebraic curves. In *2013 IEEE 10th International Symposium on Biomedical Imaging*, 860–863. IEEE.

Sun, R.; Li, Y.; Zhang, T.; Mao, Z.; Wu, F.; and Zhang, Y. 2021. Lesion-aware transformers for diabetic retinopathy grading. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10938–10947.

Sun, R.; Luo, N.; Pan, Y.; Mai, H.; Zhang, T.; Xiong, Z.; and Wu, F. 2023a. Appearance Prompt Vision Transformer for Connectome Reconstruction. In Elkind, E., ed., *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, 1423–1431. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Sun, R.; Luo, N.; Wang, Y.; Pan, Y.; Mai, H.; Zhang, Z.; and Zhang, T. 2022. 1st Place Solution for YouTubeVOS Challenge 2022: Video Object Segmentation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition Workshops*.

Sun, R.; Mai, H.; Luo, N.; Zhang, T.; Xiong, Z.; and Wu, F. 2023b. Structure-decoupled adaptive part alignment network for domain adaptive mitochondria segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 523–533. Springer.

Sun, R.; Mai, H.; Zhang, T.; and Wu, F. 2023c. DAW: Exploring the Better Weighting Function for Semi-supervised Semantic Segmentation. In *Advances in Neural Information Processing Systems*.

Sun, R.; Wang, Y.; Mai, H.; Zhang, T.; and Wu, F. 2023d. Alignment before aggregation: trajectory memory retrieval network for video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1218–1228.

Tang, Y.; Yang, D.; Li, W.; Roth, H. R.; Landman, B.; Xu, D.; Nath, V.; and Hatamizadeh, A. 2022. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20730–20740.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Vazquez-Reina, A.; Gelbart, M.; Huang, D.; Lichtman, J.; Miller, E.; and Pfister, H. 2011. Segmentation fusion for connectomics. In *2011 International Conference on Computer Vision*, 177–184. IEEE.

Wang, Y.; Luo, N.; and Zhang, T. 2023. Focus on Query: Adversarial Mining Transformer for Few-Shot Segmentation. In *Advances in Neural Information Processing Systems*.

Wang, Y.; Sun, R.; and Zhang, T. 2023. Rethinking the Correlation in Few-Shot Segmentation: A Buoys View. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7183–7192.

Wang, Y.; Sun, R.; Zhang, Z.; and Zhang, T. 2022. Adaptive agent transformer for few-shot segmentation. In *European Conference on Computer Vision*, 36–52. Springer.

Wangkai, L.; Zhaoyang, L.; Rui, S.; Huayu, M.; Naisong, L.; Wang, Y.; Yuwen, P.; Guoxin, X.; Huakai, L.; Zhiwei, X.; et al. 2023. MAUNet: Modality-Aware Anti-Ambiguity U-Net for Multi-Modality Cell Segmentation. In *Competitions in Neural Information Processing Systems*, 1–12. PMLR.

Wei, D.; Lin, Z.; Franco-Barranco, D.; Wendt, N.; Liu, X.; Yin, W.; Huang, X.; Gupta, A.; Jang, W.-D.; Wang, X.; et al. 2020. MitoEM dataset: large-scale 3D mitochondria instance segmentation from EM images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 66–76. Springer.

Xie, R.; Pang, K.; Bader, G. D.; and Wang, B. 2023. MAESTER: Masked Autoencoder Guided Segmentation at Pixel Resolution for Accurate, Self-Supervised Subcellular Structure Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3292–3301.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R. R.; and Le, Q. V. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Yuan, Z.; Yi, J.; Luo, Z.; Jia, Z.; and Peng, J. 2020. EM-net: Centerline-aware mitochondria segmentation in EM images via hierarchical view-ensemble convolutional network. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, 1219–1222. IEEE.